

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-240747

(43)Date of publication of application : 11.09.1998

(51)Int.Cl.

G06F 17/30
G06F 17/18

(21)Application number : 09-034605

(71)Applicant : INTERNATL BUSINESS MACH
CORP <IBM>

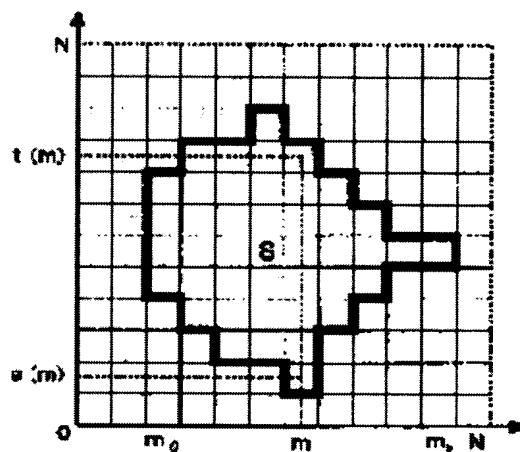
(22)Date of filing : 19.02.1997

(72)Inventor : YODA KUNIKAZU
FUKUDA TSUYOSHI
TOKUYAMA TAKESHI
MORISHITA SHINICHI(54) METHOD AND DEVICE FOR DERIVING INTER-DATA CONNECTION RULE, METHOD AND
DEVICE FOR CUTTING ORTHOGONAL PROJECTION AREA

(57)Abstract:

PROBLEM TO BE SOLVED: To make it easy to find a connection rule between data by composing a plane of two numeric attributes that an analyzed body has and cutting an orthogonal projection area between true and false attributes under specific conditions.

SOLUTION: A database having data including two kinds of numeric attribute and one kind of true/false attribute has two axes corresponding to two kinds of numeric attribute and stores the number $v(i, j)$ of data belonging to respective pixels (i row, j column) of a plane divided into $N \times N$ pixels and the number $v(i, j)$ of data whose true/false attributes are true. Then a specific condition θ is inputted to cut an orthogonal projection area S of pixels, maximizing an equation 1, out of a plane. Thus, the area in the orthogonal projection shape is cut out to make it easy for people to grasp the connection rule. Lastly, data included in the cut orthogonal projection area S are outputted.



$$\sum_{(i,j) \in S} g(i,j) = \sum_{(i,j) \in S} \{v(i,j) - \theta_{th}(i,j)\}$$

LEGAL STATUS

[Date of request for examination] 30.07.1998

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number] 3193658

[Date of registration] 25.05.2001

[Number of appeal against examiner's decision]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-240747

(43) 公開日 平成10年(1998) 9月11日

(51) Int.Cl.⁶

識別記号

F I

G 0 6 F 17/30
17/18

G 0 6 F 15/401
15/36

3 2 0 Z
Z

審査請求 未請求 請求項の数20 O L (全 31 頁)

(21) 出願番号 特願平9-34605

(22) 出願日 平成9年(1997) 2月19日

(71) 出願人 390009531

インターナショナル・ビジネス・マシー
ズ・コーポレーション

INTERNATIONAL BUSIN
ESS MASCHINES CORPO
RATION

アメリカ合衆国10504、ニューヨーク州
アーモンク (番地なし)

(72) 発明者 依田 邦和

神奈川県大和市下鶴間1623番地14 日本ア
イ・ビー・エム株式会社 東京基礎研究所
内

(74) 代理人 弁理士 合田 潔 (外2名)

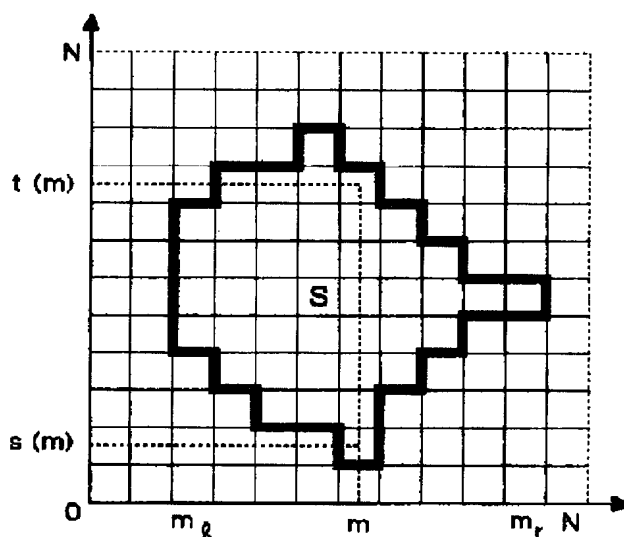
最終頁に続く

(54) 【発明の名称】 データ間結合ルール導出方法及び装置、及び直交凸領域切出方法及び装置

(57) 【要約】

【課題】 2項の数値属性と真偽をとる属性を有するデータ間の結合ルールを人間がより把握しやすい形で提示すること。

【解決手段】 (1) 2つの数値属性により平面を構成し、この平面をピクセルに分割し、各ピクセル内のデータ数及び真偽をとる属性が真となったデータの数のカウントする。(2) 所定の条件 θ に従い、平面の2つの軸に凸な領域である直交凸領域(rectilinear region)を切り出し、データ間の結合ルールを見い出す。(3) 切り出した直交凸領域が、先に述べたようなサポート最大化ルール等の条件を満たしていれば、その直交凸領域をユーザに提示する。また、データベースからその直交凸領域に含まれるデータの必要な属性を引き出すことも必要に応じて行う。



【特許請求の範囲】

【請求項 1】 2 種類の数値属性と、 1 種類の真偽をとる属性とを含むデータを有するデータベースにおいて、データ間の結合ルールを導き出す方法であって、前記 2 種類の数値属性に対応する 2 つの軸を有し且つ $N \times M$ 個のピクセルに分割されている平面の各ピクセルに対応して、当該ピクセル (i 行 j 列) に属するデータの数 $u(i, j)$ 及び前記真偽をとる属性が真であるデータの数 $v(i, j)$ を記憶する平面構成ステップと、条件 θ を入力するステップと、

【数 1】

$$\sum_{(i,j) \in S} g(i, j) = \sum_{(i,j) \in S} (v(i, j) - \theta u(i, j))$$

を最大にするような前記ピクセルの直交凸領域 S を前記平面から切り出す領域切出ステップと、切り出された前記直交凸領域 S 内に含まれるデータを出力するステップとを含むデータ間結合ルール導出方法。

【請求項 2】 入力された前記条件 θ とは異なる第 2 の条件 θ_2 を入力するステップと、

【数 2】

$$\sum_{(i,j) \in S_2} g(i, j) = \sum_{(i,j) \in S_2} (v(i, j) - \theta_2 u(i, j))$$

を最大にするような前記ピクセルの第 2 の直交凸領域 S_2 を前記平面から切り出すステップと、

【数 3】

$$\theta_3 = \frac{V(S_2) - V(S)}{U(S_2) - U(S)}$$

(前記直交凸領域 S_2 に含まれ且つ前記真偽をとる属性が真であるデータの数を $V(S_2)$ 、前記直交凸領域 S に含まれ且つ前記真偽をとる属性が真であるデータの数を $V(S)$ 、前記直交凸領域 S_2 に含まれるデータ数を $U(S_2)$ 、前記直交凸領域 S に含まれるデータ数を $U(S)$ とする。) を第 3 の条件として、

【数 4】

$$\sum_{(i,j) \in S_3} g(i, j) = \sum_{(i,j) \in S_3} (v(i, j) - \theta_3 u(i, j))$$

を最大にするような前記ピクセルの第 3 の直交凸領域 S_3 を前記平面から切り出すステップとをさらに含む請求項 1 記載のデータ間結合ルール導出方法。

【請求項 3】 前記切り出された直交凸領域 S 内の各ピクセルの $v(i, j)$ 、 $u(i, j)$ が、前記平面全体のデータ数に対する前記平面全体の前記真偽をとる属性が真であるデータ数の割合に等しくなるよう $v(i, j)$ を変更するステップと、当該変更された $v(i, j)$ を用いて、入力された条件 θ_4 に従い、

【数 5】

$$\sum_{(i,j) \in S_4} g(i, j) = \sum_{(i,j) \in S_4} (v(i, j) - \theta_4 u(i, j))$$

を最大にするような前記ピクセルの第 4 の直交凸領域 S_4 を切り出すステップとをさらに含む請求項 1 記載のデータ間結合ルール導出方法。

【請求項 4】 前記平面構成ステップが、複数の前記データから、 X 個のデータをランダムサンプリングするステップと、

- 10 サンプリングされたデータを各前記数値属性についてソートし、 $X \cdot i \leq N$ ($i = 1, 2, \dots, N$) 番目に該当する数値及び $X \cdot n \leq M$ ($n = 1, 2, \dots, M$) 番目に該当する数値を記憶するステップと、記憶された前記数値を基準にして、前記複数のデータの各々が $N \times M$ 個の前記ピクセルのいずれに含まれるか判断し、各ピクセルにおける数を計数するステップとを含む請求項 1 記載のデータ間結合ルール導出方法。

【請求項 5】 前記領域切出ステップが、

- 第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が前記区間 $[s, t]$ に含まれる直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 1 の値が最も大きい直交凸領域 $S_1^m(s, t)$ の前記第 $m-1$ 列の区間 $[x, y]$ 又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^1 に記憶し、当該直交凸領域 $S_1^m(s, t)$ の数 1 の値を記憶する第 1 記憶ステップと、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $s \geq x$ 及び $t \geq y$ を満たす直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 1 の値が最も大きい直交凸領域 $S_2^m(s, t)$ の (a) 前記第 $m-1$ 列の区間 $[x, y]$ 及び (b) 前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^2 に記憶し、当該直交凸領域 $S_2^m(s, t)$ の数 1 の値を記憶する第 2 記憶ステップと、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $s \leq x$ 及び $y \geq t$ を満たす直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 1 の値が最も大きい直交凸領域 $S_3^m(s, t)$ の (a) 前記第 $m-1$ 列の区間 $[x, y]$ 及び (b) 前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^3 に記憶し、当該直交凸領域 $S_3^m(s, t)$ の数 1 の値を記憶する第 3 記憶ステップと、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $x \leq s$ 及び $y \geq t$ を満たす直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 1 の値が最も大きい直交凸領域

域 $S_m^N(s, t)$ の (a) 前記第 $m-1$ 列の区間 $[x, y]$ 及び (b) 前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S_m^N(s, t)$ の数 1 の値を記憶する第 4 記憶ステップと、

全ての m 及び $[s, t]$ について前記第 1 乃至第 4 記憶ステップを実行し、最も大きい数 1 の値を有する直交凸領域 S の m 及び $[s, t]$ と、対応する記憶手段 H^N 、 H^D 、又は H^N の値とを用いて、直交凸領域 S を前記平面から切り出すステップとを含む請求項 1 記載のデータ間結合ルール導出方法。

【請求項 6】 各々内部に含まれるポイントの数 $u(i, j)$ 及び所定の条件を満たしたポイントの数 $v(i, j)$ を記憶した複数のセルを含む平面から、

$$\text{【数 6】} \quad \sum_{(i, j) \in S} g(i, j) = \sum_{(i, j) \in S} (v(i, j) - \theta u(i, j))$$

を最大とする直交凸領域 S を切り出す方法であって、 θ を入力するステップと、

第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が前記区間 $[s, t]$ に含まれる直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 6 の値が最も大きい直交凸領域 $S_m^N(s, t)$ の前記第 $m-1$ 列の区間 $[x, y]$ 又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S_m^N(s, t)$ の数 6 の値を記憶する第 1 記憶ステップと、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $s \geq x$ 及び $t \geq y$ を満たす直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 6 の値が最も大きい直交凸領域 $S_m^N(s, t)$ の (a) 前記第 $m-1$ 列の区間 $[x, y]$ 及び (b) 前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S_m^N(s, t)$ の数 6 の値を記憶する第 2 記憶ステップと、

第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $s \leq x$ 及び $y \geq t$ を満たす直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 6 の値が最も大きい直交凸領域 $S_m^D(s, t)$ の (a) 前記第 $m-1$ 列の区間 $[x, y]$ 及び (b) 前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^D に記憶し、当該直交凸領域 $S_m^D(s, t)$ の数 6 の値を記憶する第 3 記憶ステップと、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $x \leq s$ 及び $y \geq t$ を満たす直交凸領

域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 6 の値が最も大きい直交凸領域 $S_m^N(s, t)$ の (a) 前記第 $m-1$ 列の区間 $[x, y]$ 及び (b) 前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S_m^N(s, t)$ の数 6 の値を記憶する第 4 記憶ステップと、

全ての m 及び $[s, t]$ について前記第 1 乃至第 4 記憶ステップを実行し、最も大きい数 6 の値を有する直交凸領域 S の m 及び $[s, t]$ と、対応する記憶手段 H^N 、 H^D 、又は H^N の値とを用いて、直交凸領域 S を前記平面から切り出すステップとを含む直交凸領域切出方法。

【請求項 7】 前記第 2 記憶ステップが、第 m 列の区間 $[s, t]$ が右端列であり且つ第 $m-1$ 列の区間 $[x, y]$ が $s \geq x$ 及び $t = y$ を満たし且つ第 $m-2$ 列の区間 $[a, b]$ と前記第 $m-1$ 列の区間 $[x, y]$ との関係が $a \geq x$ 及び $b \leq y$ 又は $a \leq x$ 及び $b \leq y$ である直交凸領域、第 m 列の区間 $[s, t]$ が右端列であり且つ第 $m-1$ 列の区間 $[x, y]$ が $s \geq x$ 及び $s \leq y \leq t-1$ を満たし且つ第 $m-2$ 列の区間 $[a, b]$ と前記第 $m-1$ 列の区間 $[x, y]$ との関係が $a \geq x$ 及び $b \leq y$ 又は $a \leq x$ 及び $b \leq y$ である直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 6 の値が最も大きい直交凸領域 $S_m^N(s, t)$ の (a) 前記第 $m-1$ 列の区間 $[x, y]$ 及び (b) 前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S_m^N(s, t)$ の数 6 の値を記憶することを特徴とする請求項 6 記載の直交凸領域切出方法。

【請求項 8】 前記第 3 記憶ステップが、第 m 列の区間 $[s, t]$ が右端列であり且つ第 $m-1$ 列の区間 $[x, y]$ が $s = x$ 及び $t \geq y$ を満たし且つ第 $m-2$ 列の区間 $[a, b]$ と前記第 $m-1$ 列の区間 $[x, y]$ との関係が $a \geq x$ 及び $b \leq y$ 又は $a \geq x$ 及び $b \geq y$ である直交凸領域、第 m 列の区間 $[s, t]$ が右端列であり且つ第 $m-1$ 列の区間 $[x, y]$ が $s+1 \leq x \leq t$ 及び $y \geq t$ を満たし且つ第 $m-2$ 列の区間 $[a, b]$ と前記第 $m-1$ 列の区間 $[x, y]$ との関係が $a \geq x$ 及び $b \leq y$ 又は $a \geq x$ 及び $b \geq y$ である直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 6 の値が最も大きい直交凸領域 $S_m^D(s, t)$ の (a) 前記第 $m-1$ 列の区間 $[x, y]$ 及び (b) 前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^D に記憶し、当該直交凸領域 $S_m^D(s, t)$ の数 6 の値を記憶す

ることを特徴とする請求項6記載の直交凸領域切出方法。

【請求項9】前記第4記憶ステップが、第m列の区間〔s, t〕が右端列であって第m-1列の区間〔x, y〕が前記区間〔s, t〕と同一である直交凸領域、第m列の区間〔s, t〕が右端列であって第m-1列の区間〔x, y〕が $x \leq s$ 及び $y \geq t+1$ 又は $x \leq s-1$ 及び $y \geq t$ を満たす直交凸領域、又は前記第m列の区間〔s, t〕のみで構成される直交凸領域のうち、前記数6の値が最も大きい直交凸領域 $S_m^N(s, t)$ の

(a)前記第m-1列の区間〔x, y〕及び(b)前記第m-1列の区間〔x, y〕と第m-2列の区間〔a, b〕との関係、又は第m列が左端列であることを示す情報をm及び〔s, t〕に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S_m^N(s, t)$ の数6の値を記憶することを特徴とする請求項6記載の直交凸領域切出方法。

【請求項10】前記直交凸領域Sを前記平面から切り出すステップが、全てのm及び〔s, t〕について前記第1乃至第4記憶ステップを実行するステップと、

計算された数6の値のうち最大の値を有する直交凸領域Sのm及び〔s, t〕と、対応する記憶手段 H^N 、 H^D 、 H^D 、又は H^N を用いて、第m-1列の区間〔x, y〕及び前記第m-1列と第m-2列との関係を読み出す第1関係読出ステップと、

前記第m-1列と第m-2列との関係を用いて、前記記憶手段 H^N 、 H^D 、 H^D 、及び H^N から対応する記憶手段を選択する選択ステップと、

前記第m-1列の区間〔x, y〕を用いて選択された記憶手段から第m-2列の区間〔a, b〕及び第m-2列と第m-3列との関係を読み出す第2関係読出ステップと、

前記選択ステップと前記第2関係読出ステップとを、前列との関係が前記左端列であることを示す情報となるまで繰り返すステップとを含む請求項6記載の直交凸領域切出方法。

【請求項11】2種類の数値属性と、1種類の真偽をとる属性を含むデータを有するデータベースにおいて、データ間の結合ルールを導き出す装置であって、前記2種類の数値属性に対応する2つの軸を有し且つ $N \times M$ 個のピクセルに分割されている平面の各ピクセルに対応して、当該ピクセル(i行j列)に属するデータの数 $u(i, j)$ 及び前記真偽をとる属性が真であるデータの数 $v(i, j)$ を記憶する平面構成装置と、条件 θ を入力する入力デバイスと、

【数7】
$$\sum_{(i,j) \in S} g(i, j) = \sum_{(i,j) \in S} (v(i, j) - \theta u(i, j))$$

を最大にするような前記ピクセルの直交凸領域Sを前記平面から切り出す領域切出装置と、

(4)

特開平10-240747

6

切り出された前記領域S内に含まれるデータを出力するデバイスとを有するデータ間結合ルール導出装置。

【請求項12】前記入力デバイスにより、前記条件 θ とは異なる第2の条件 θ_2 を入力し、前記領域切出装置により、前記第2の条件 θ_2 に対応する第2の直交凸領域 S_2 を前記平面から切り出した場合に、

【数8】

$$\theta_3 = \frac{V(S_2) - V(S)}{U(S_2) - U(S)}$$

(前記直交凸領域 S_2 に含まれ且つ前記真偽をとる属性が真であるデータの数を $V(S_2)$ 、前記直交凸領域Sに含まれ且つ前記真偽をとる属性が真であるデータの数を $V(S)$ 、前記直交凸領域 S_2 に含まれるデータ数を $U(S_2)$ 、前記直交凸領域Sに含まれるデータ数を $U(S)$ とする。)を第3の条件として前記領域切出装置に出力する手段とをさらに有する請求項11記載のデータ間結合ルール導出装置。

【請求項13】前記切り出された直交凸領域S内の各ピクセルの $v(i, j)$ 及び $u(i, j)$ が、前記平面全体のデータ数に対する前記平面全体の前記真偽をとる属性が真であるデータ数の割合に等しくなるよう $v(i, j)$ を変更する手段と、

当該変更された $v(i, j)$ 及び入力された条件 θ_4 でもって、前記領域切出装置が動作するように命令する手段とを有する請求項11記載のデータ間結合ルール導出装置。

【請求項14】前記平面構成装置が、複数の前記データから、X個のデータをランダムサンプリングする手段と、

サンプリングされたデータを各前記数値属性についてソートし、 $X \cdot i \div N$ ($i = 1, 2, \dots, N$) 番目に該当する数値及び $X \cdot n \div M$ ($n = 1, 2, \dots, M$) 番目に該当する数値を記憶する手段と、

記憶された前記数値を基準にして、前記複数のデータの各々が $N \times M$ 個の前記ピクセルのいずれに含まれるか判断し、各ピクセルにおける数を計数する手段とを含む請求項11記載のデータ間結合ルール導出装置。

【請求項15】各々内部に含まれるポイントの数 $u(i, j)$ 及び所定の条件を満たしたポイントの数 $v(i, j)$ を記憶した複数のセルを含む平面から、

【数9】

$$\sum_{(i,j) \in S} g(i, j) = \sum_{(i,j) \in S} (v(i, j) - \theta u(i, j))$$

を最大とする直交凸領域Sを切り出す装置であって、 θ を入力する手段と、

第m列の区間〔s, t〕が右端列であって第m-1列の区間〔x, y〕が前記区間〔s, t〕に含まれる直交凸領域、又は前記第m列の区間〔s, t〕のみで構成される直交凸領域のうち、前記数9の値が最も大きい直交凸

50

領域 $S_m^N(s, t)$ の前記第 $m-1$ 列の区間 $[x, y]$ 又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S_m^N(s, t)$ の数 9 の値を記憶する第 1 記憶手段と、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $s \geq x$ 及び $t \geq y$ を満たす直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 9 の値が最も大きい直交凸領域 $S_m^N(s, t)$ の (a) 前記第 $m-1$ 列の区間 $[x, y]$ 及び (b) 前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S_m^N(s, t)$ の数 9 の値を記憶する第 2 記憶手段と、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $s \leq x$ 及び $y \geq t$ を満たす直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 9 の値が最も大きい直交凸領域 $S_m^N(s, t)$ の (a) 前記第 $m-1$ 列の区間 $[x, y]$ 及び (b) 前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S_m^N(s, t)$ の数 9 の値を記憶する第 3 記憶手段と、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $x \leq s$ 及び $y \geq t$ を満たす直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数 9 の値が最も大きい直交凸領域 $S_m^N(s, t)$ の (a) 前記第 $m-1$ 列の区間 $[x, y]$ 及び (b) 前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S_m^N(s, t)$ の数 9 の値を記憶する第 4 記憶手段と、全ての m 及び $[s, t]$ について前記第 1 乃至第 4 記憶手段を動作させ、最も大きい数 9 の値を有する直交凸領域 S の m 及び $[s, t]$ と、対応する記憶手段 H^N 、 H^N 、 H^N 、又は H^N の値とを用いて、直交凸領域 S を前記平面から切り出す手段とを有する直交凸領域切出装置。

【請求項 16】2 種類の数値属性と、1 種類の真偽をとる属性を含むデータを有するデータベースにおいて、コンピュータにデータ間の結合ルールを導き出させるプログラムを記憶した記憶デバイスであって、前記プログラムは、前記 2 種類の数値属性に対応する 2 つの軸を有し且つ $N \times M$ 個のバケットに分割されている平面の各ピクセルに対応して、当該ピクセル (i 行 j 列) に属するデータの数 $u(i, j)$ 及び前記真偽をとる属性が真であるデータの数 $v(i, j)$ を記憶する平面構成ステップと、

(5)

特開平 10-240747

8

条件 θ を入力する入力ステップと、

【数 10】

$$\sum_{(i,j) \in S} g(i, j) = \sum_{(i,j) \in S} (v(i, j) - \theta u(i, j))$$

を最大にするような前記ピクセルの直交凸領域 S を前記平面から切り出す領域切出ステップとをコンピュータに実行させる、記憶デバイス。

【請求項 17】前記プログラムが、前記条件 θ とは異なる第 2 の条件 θ_2 を入力するステップと、

【数 11】

$$\sum_{(i,j) \in S_2} g(i, j) = \sum_{(i,j) \in S_2} (v(i, j) - \theta_2 u(i, j))$$

を最大とするような前記ピクセルの第 2 の直交凸領域 S_2 を前記平面から切り出すステップと、

【数 12】

$$\theta_3 = \frac{V(S_2) - V(S)}{U(S_2) - U(S)}$$

20 (前記直交凸領域 S_2 に含まれ且つ前記真偽をとる属性が真であるデータの数を $V(S_2)$ 、前記直交凸領域 S に含まれ且つ前記真偽をとる属性が真であるデータの数を $V(S)$ 、前記直交凸領域 S_2 に含まれるデータ数を $U(S_2)$ 、前記直交凸領域 S に含まれるデータ数を $U(S)$ とする。) を第 3 の条件として、

【数 13】

$$\sum_{(i,j) \in S_3} g(i, j) = \sum_{(i,j) \in S_3} (v(i, j) - \theta_3 u(i, j))$$

30

を最大にするような前記ピクセルの第 3 の直交凸領域 S_3 を前記平面から切り出すステップとをコンピュータに実行させる、請求項 16 記載の記憶デバイス。

【請求項 18】前記プログラムが、前記切り出された直交凸領域 S 内の各ピクセルの $v(i, j)$ 、 $u(i, j)$ が、前記平面全体のデータ数に対する前記平面全体の前記真偽をとる属性のデータ数の割合に等しくなるよう $v(i, j)$ を変更するステップと、当該変更された $v(i, j)$ 及び入力された条件 θ_4 に従い、

【数 14】

$$\sum_{(i,j) \in S_4} g(i, j) = \sum_{(i,j) \in S_4} (v(i, j) - \theta_4 u(i, j))$$

40

を最大にするような前記ピクセルの第 4 の直交凸領域 S_4 を切り出すステップとをコンピュータに実行させる、請求項 16 記載の記憶デバイス。

【請求項 19】前記平面構成ステップが、複数の前記データから、 X 個のデータをランダムサンプリングするステップと、サンプリングされたデータを各前記数値属性についてソートし、 $X \cdot i \div N$ ($i = 1, 2, \dots, N$) 番目に該当する数値及び $X \cdot n \div M$ ($n = 1, 2, \dots, M$) 番目に

50

該当する数値を記憶するステップと、記憶された前記数値を基準にして、前記複数のデータの各々が $N \times M$ 個の前記ピクセルのいずれに含まれるか判断し、各ピクセルにおける数を計数するステップとを含む請求項16記載の記憶デバイス。

【請求項20】各々内部に含まれるポイントの数 $u(i, j)$ 及び所定の条件を満たしたポイントの数 $v(i, j)$ を記憶した複数のセルを含む平面から、

【数15】

$$\sum_{(i,j) \in S} g(i, j) = \sum_{(i,j) \in S} (v(i, j) - \theta u(i, j))$$

を最大とする直交凸領域 S を切り出すプログラムを格納した記憶媒体であって、

前記プログラムは、

θ を入力するステップと、

第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が前記区間 $[s, t]$ に含まれる直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数15の値が最も大きい直交凸領域 $S_m^*(s, t)$ の前記第 $m-1$ 列の区間 $[x, y]$ 又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^* に記憶し、当該直交凸領域 $S_m^*(s, t)$ の数15の値を記憶する第1記憶ステップと、

第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $s \geq x$ 及び $t \geq y$ を満たす直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数15の値が最も大きい直交凸領域 $S_m^*(s, t)$ の(a)前記第 $m-1$ 列の区間 $[x, y]$ 及び(b)前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^* に記憶し、当該直交凸領域 $S_m^*(s, t)$ の数15の値を記憶する第2記憶ステップと、

第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $s \leq x$ 及び $y \geq t$ を満たす直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数15の値が最も大きい直交凸領域 $S_m^0(s, t)$ の(a)前記第 $m-1$ 列の区間 $[x, y]$ 及び(b)前記第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^0 に記憶し、当該直交凸領域 $S_m^0(s, t)$ の数15の値を記憶する第3記憶ステップと、

第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $x \leq s$ 及び $y \geq t$ を満たす直交凸領域、又は前記第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、前記数15の値が最も大きい直交凸領域 $S_m^*(s, t)$ の(a)前記第 $m-1$ 列の区間 $[x, y]$ 及び(b)前記第 $m-1$ 列の区間 $[x, y]$ と第 m

−2列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^* に記憶し、当該直交凸領域 $S_m^*(s, t)$ の数15の値を記憶する第4記憶ステップと、

全ての m 及び $[s, t]$ について前記第1乃至第4記憶ステップを実行し、最も大きい数15の値を有する直交凸領域 S の m 及び $[s, t]$ と、対応する記憶手段 H^* 、 H^0 、 H^0 、又は H^* の値とを用いて、直交凸領域 S を前記平面から切り出すステップとをコンピュータに実行させる、記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、データベースにおけるデータ相関の解析（データマイニングという。）に関し、より詳しくは2項の数値属性と1項の真偽をとる属性（真偽をとる条件又は0-1属性ともいう。）を有するデータ間の相関を見出す手法に関する。

【0002】

【従来の技術】例えば、銀行の顧客を解析対象とし、流動性預金残高がいくらからいで且つ年齢が何歳ぐらいの人であれば、定期預金残高が200万円以上になる人が全体の20%となるか、といった問題を実際に解くことを考える。この流通性預金残高及び年齢は、整数ではあるが連続数値であり、一方定期預金残高200万円以上というのは、200万円以上か未満かという分類になるので、真偽をとる属性を有するものである。真偽をとる属性は、例えば「顧客がクレジットカードを有しているか」や「顧客が男性であるか」といった問題と置き換えることも可能である。このような課題を解決することができれば、銀行はどのような人に、例えば新型の金融商品に関するダイレクトメールを送ればよいか簡単に分かるので、効率的な営業活動が行える。

【0003】従来、先に述べた真偽をとる属性間の相関を表現するルール（結合ルール、association rule）を高速に抽出するような研究は、データマイニングの分野において行われてきた。例えば、R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases" In proceedings of the ACM SIGMOD Conference on Management of data, May 1993. や、R. Agrawal and R. Srikant, "Fast algorithms for mining association rules" In Proceedings of the 20th VLDB Conference, 1994. 等がある。

【0004】また、2項の数値データ間のルールを求める従来手法には、以下のようなものがある。

1. 強い線形相関を見出すために、平面上の直線で、点集合を最適近似するものを探す方法。例えば、最小白乗法、再帰中央法等である。これら方法の欠点は、線形相関しか分からず、しかも相関係数の絶対値が0.5以下の場合に線形相関を用いて各データを予測すると精度が低く、現実にはほとんど役に立たない点にある。

2. 弱い大域相関を見出すために、2次元平面上で正方形、長方形、又は円、楕円で面積に対して多くのデータを含むものを見出す方法。例えば、計算幾何学アルゴリズムを利用するものである。この場合、計算時間が大きくなってしまいうという欠点がある。例えば円の場合、 $O(M^3)$ 以上の手間が掛かり得る($O(M^3)$ は、オーダー M^3 の計算時間がかかることを示す。 M はデータ数である。)。また、取り出す相関領域としては決まった形をしたものしか扱うことができない。現実には、決まった形で適切にカバーできる場合は少ない。

3. 平面を正方メッシュに分割しておき、たくさんのデータを含むピクセルを取り出す方法。しかし、取り出されたピクセルの集合は連結でなく、バラバラなことが多いので、ルールとして見出すのは困難である。

【0005】このような手法を用いると、上記の欠点の他に、データ間の多くのルールのうちで、意味のあるものと無意味なものとの区別が難しいという欠点もある。通常、相関に実用上の意味があるかどうかは人間の判断によらないといけなことが多いが、1. や2. では特殊な相関しか取り出せないで意味ある相関を見逃しやすく、3では出力を人間が見てルールを見い出せない。

【0006】他の方法としては、平面を正方メッシュに分割しておき、これらのピクセルに関して連結且つ x 単調な領域のうち多くのデータを含む領域を切り出す方法がある(Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita and Takeshi Tokuyama, "Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization," In Proceedings of the ACM SIGMOD Conference on Management of Data, pages 13-23, June 1996 を参照のこと)。 x 単調とは、列方向には凸であるが、行方向では凸ではないものを言う。この方法は、高速で、一定の意味ある相関を取り出すことができるが、縦方向に激しく揺れる入り組んだ領域を切り出すことが多く、人間が見てどこが強い相関の部分であるか把握ににくい。また、 x 単調ということで、切り出される領域の形状が、正方メッシュのメッシュの仕方(各ピクセルへのデータの配分の仕方)に大きく依存するという欠点もある。

【0007】

【発明が解決しようとする課題】本発明は、以上のような点に鑑み、2項以上の数値属性と真偽をとる属性を有するデータ間の結合ルールを見出すための一手法を提供することを目的とする。

【0008】また、データ間の結合ルールを人間がより把握しやすい形で提示することも目的である。そして、多くの結合ルールを可視化することにより、使用する人間の選択の幅を増大させ、より重要な結合ルールを見いだすこと可能とすることも目的とする。

【0009】また、(1)真偽をとる属性が真であるデータの割合がある定められた値以上であって、含まれる

データ数が最大となるようなルールであるサポート最大化ルールや、(2)最低限含まれるデータ数が定められた場合、真偽をとる属性が真であるデータの割合が最大となるようなルールであるコンフィデンス最大化ルール、(3)取り出される領域内部と外部との分割を考えた時に、分割前の情報量と比較した分割後の情報量の増分を最大化するルールである最適化エントロピー・ルール、(4)領域内外の分割を考えた時に、内外の「標準化された真偽の割合の平均からのずれ」の二乗和を最大化するルールである最適化インタクラスバリエーション・ルールを満たすような範囲(領域)を導出可能とすることも目的である。

【0010】さらに、上記のようなデータ間の結合ルールを高速に実行できるような手法を提供することも目的である。

【0011】

【課題を解決するための手段】通常、解析対象物は多くの数値属性を有する。この中から2つの数値属性を選び、また、1つの真偽をとる属性について、以下のステップを行うことにより、上記の目的を達成するものである。すなわち、

(1) 2つの数値属性により平面を構成し、この平面をピクセルに分割し、各ピクセル内のデータ数及び真偽をとる属性が真となったデータの数をカウントする。このような平面は、データ数が濃淡度、真偽をとる属性が真となるデータの数が彩度に該当するような、複数のピクセルを有する平面画像として捉えることもできる。

(2) 所定の条件 θ に従い、平面の2つの軸に凸な領域である直交凸領域(rectilinear region)を切り出し、データ間の結合ルールを見出す。

(3) 切り出した直交凸領域が、先に述べたようなサポート最大化ルール等の条件を満たしていれば、その直交凸領域をユーザに提示する。また、データベースからその直交凸領域に含まれるデータの必要な属性を引き出すことも、必要に応じて行う。

【0012】なお、切り出された直交凸領域を、そのままユーザに提示したり、複数の直交凸領域を切り出した場合には、それを動画として可視化することにより、所望の結合ルールを見出し易くすることもできる。

【0013】また、一旦直交凸領域を切り出した後に、それ以外の結合ルールを見出すべく、切り出された直交凸領域について、彩度を平均化し、再度切り出しステップを実行することも可能である。

【0014】最初に述べたような例の場合、流動性預金残高の軸と、年齢の軸を設け、その平面を適当なメッシュに分割する。そして、メッシュの各エレメントであるピクセルについて該当する顧客の数と、定期預金残高200万円以上の顧客の数をカウントする。そして、例えば顧客全体の20%が入り且つ定期預金残高200万円以上である顧客の割合が最大となるような直交凸領域で

ある領域の切り出しを行うことにより、コンフィデンス最大化ルールを得ることができる。

【0015】また、例えば定期預金残高200万円以上の顧客割合が10%で最大の顧客数を有する直交凸領域を切り出すことにより、サポート最大化ルールを得ることができる。

【0016】以上述べた事項をまとめると、2種類の数値属性と、1種類の真偽をとる属性とを含むデータを有するデータベースにおいて、まず、2種類の数値属性に対応する2つの軸を有し且つ $N \times M$ 個のピクセルに分割されている平面の各ピクセルに対応して、当該ピクセル（ i 行 j 列）に属するデータの数 $u(i, j)$ 及び前記真偽をとる属性が真であるデータの数 $v(i, j)$ を記憶する（平面構成ステップ）。次に、所定の条件 θ を入力する。そして、

【数16】

$$\sum_{(i,j) \in S} g(i, j) = \sum_{(i,j) \in S} (v(i, j) - \theta u(i, j))$$

を最大にするような、ピクセルの直交凸領域 S を平面から切り出す（領域切出ステップ）。このように直交凸な形状の領域を切り出すことにより、より人間に結合ルールが把握しやすいようになる。また、先の平面構成ステップからの依存性が小さくできる。最後に、切り出された直交凸領域 S 内に含まれるデータを出力する。このように、ルールに合致するデータを得ることができる。

【0017】また、入力された条件 θ とは異なる第2の条件 θ_2 を入力し、

【数17】

$$\sum_{(i,j) \in S_2} g(i, j) = \sum_{(i,j) \in S_2} (v(i, j) - \theta_2 u(i, j))$$

を最大にするようなピクセルの第2の直交凸領域 S_2 を平面から切り出し、さらに、

【数18】

$$\theta_3 = \frac{V(S_2) - V(S)}{U(S_2) - U(S)}$$

（直交凸領域 S_2 に含まれ且つ真偽をとる属性が真であるデータの数を $V(S_2)$ 、直交凸領域 S に含まれ且つ真偽をとる属性が真であるデータの数を $V(S)$ 、直交凸領域 S_2 に含まれるデータ数を $U(S_2)$ 、直交凸領域 S に含まれるデータ数を $U(S)$ とする。）を第3の条件として、

【数19】

$$\sum_{(i,j) \in S_3} g(i, j) = \sum_{(i,j) \in S_3} (v(i, j) - \theta_3 u(i, j))$$

を最大にするようなピクセルの第3の領域 S_3 を平面から切り出すようにすることも考えられる。このような処理は、最初の条件 θ で、初期の目的のルールを導き出せなかった場合に有用である。通常先に示したサポート最大化ルール、コンフィデンス最大化ルール、最適化エン

トロピ・ルール、最適化インタクラスバリエーション・ルールといったルールを求める際には、条件 θ を適当に変化させ、上記のような処理を行うことにより求められる。

【0018】さらに、切り出された直交凸領域 S 内の各ピクセルの $v(i, j)$ と $u(i, j)$ が、平面全体のデータ数に対する平面全体の真偽をとる属性が真であるデータ数の割合に等しくなるよう $v(i, j)$ を変更し、当該変更された $v(i, j)$ を用いて、入力された条件 θ_4 に従い、

【数20】

$$\sum_{(i,j) \in S_4} g(i, j) = \sum_{(i,j) \in S_4} (v(i, j) - \theta_4 u(i, j))$$

を最大にするようなバケットの第4の領域 S_4 を切り出すようにすることも考えられる。このようにすると、二次的な相関ルールを導き出すことができる。

【0019】また、先の平面構成ステップは、複数のデータから、 X 個のデータをランダムサンプリングするステップと、サンプリングされたデータを各数値属性についてソートし、 $X \cdot i \leq N$ （ $i = 1, 2, \dots, N$ ）番目に該当する数値及び $X \cdot n \leq M$ （ $n = 1, 2, \dots, M$ ）番目に該当する数値を記憶し、記憶された数値を基準にして、複数のデータを $N \times M$ 個のピクセルに入れるようにすることも考えられる。このようにすると、各行各列にデータを高速にまたほぼ均等に割り振ることができる。

【0020】領域切出ステップは本発明の主要部分である。ここで、第 m 列の区間 $[s, t]$ が右端列であるような直交凸領域は、第 $m-1$ 列から第 m 列に移行する際に、第 $m-1$ 列の区間 $[x, y]$ に比して、（1）広がるか、（2）上昇するか、（3）下降するか、（4）狭まるかの4つの類型に分けられる。そして、4つの類型のうち、最も大きい数16の値を有する直交凸領域が求めたい領域である。よって、それぞれの類型について最大の数16の値を有する領域を求めるため、以下のようなステップが実行される。

【0021】第1の類型のため、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が区間 $[s, t]$ に含まれる直交凸領域、又は第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、数16の値が最も大きい直交凸領域 $S_a^N(s, t)$ の第 $m-1$ 列の区間 $[x, y]$ 又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S_a^N(s, t)$ の数16の値を記憶する。第2の類型のために、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $s \geq x$ 及び $t \geq y$ を満たす直交凸領域、又は第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、数16の値が最も大きい直交凸領域 $S_b^N(s, t)$ の（a）第 $m-1$ 列の区間 $[x, y]$ 及び（b）第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応

して記憶手段 H^N に記憶し、当該直交凸領域 $S^N(s, t)$ の数16の値を記憶する。

【0022】第3の類型のため、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $s \leq x$ 及び $y \geq t$ を満たす直交凸領域、又は第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、数16の値が最も大きい直交凸領域 $S^N(s, t)$ の(a)第 $m-1$ 列の区間 $[x, y]$ 及び(b)第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S^N(s, t)$ の数16の値を記憶する。最後に、第4の類型のため、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $x \leq s$ 及び $y \geq t$ を満たす直交凸領域、又は第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、数16の値が最も大きい直交凸領域 $S^N(s, t)$ の(a)第 $m-1$ 列の区間 $[x, y]$ 及び(b)第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S^N(s, t)$ の数16の値を記憶する。

【0023】各々のステップは、直交凸という性質を考慮して構成されている。そして、全ての m 及び $[s, t]$ について前記第1乃至第4記憶ステップを実行し、最も大きい数16の値を有する直交凸領域 S の m 及び $[s, t]$ と、対応する記憶手段 H^N 、 H^N 、 H^N 、又は H^N の値とを用いて、直交凸領域 S を平面から切り出す。

【0024】なお、第2の類型のための計算は、より詳しく説明すると、第 m 列の区間 $[s, t]$ が右端列であり且つ第 $m-1$ 列の区間 $[x, y]$ が $s \geq x$ 及び $t = y$ を満たし且つ第 $m-2$ 列の区間 $[a, b]$ と第 $m-1$ 列の区間 $[x, y]$ との関係が $a \geq x$ 及び $b \leq y$ 又は $a \leq x$ 及び $b \geq y$ である直交凸領域、第 m 列の区間 $[s, t]$ が右端列であり且つ第 $m-1$ 列の区間 $[x, y]$ が $s \geq x$ 及び $s \leq y \leq t-1$ を満たし且つ第 $m-2$ 列の区間 $[a, b]$ と第 $m-1$ 列の区間 $[x, y]$ との関係が $a \geq x$ 及び $b \leq y$ 又は $a \leq x$ 及び $b \geq y$ である直交凸領域、又は第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、数16の値が最も大きい直交凸領域 $S^N(s, t)$ の(a)第 $m-1$ 列の区間 $[x, y]$ 及び(b)第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S^N(s, t)$ の数16の値を記憶する、という処理になる。

【0025】また、第3の類型のための計算は、より詳しく説明すると、第 m 列の区間 $[s, t]$ が右端列であり且つ第 $m-1$ 列の区間 $[x, y]$ が $s = x$ 及び $t \geq y$

を満たし且つ第 $m-2$ 列の区間 $[a, b]$ と第 $m-1$ 列の区間 $[x, y]$ との関係が $a \geq x$ 及び $b \leq y$ 又は $a \leq x$ 及び $b \geq y$ である直交凸領域、第 m 列の区間 $[s, t]$ が右端列であり且つ第 $m-1$ 列の区間 $[x, y]$ が $s+1 \leq x \leq t$ 及び $y \geq t$ を満たし且つ第 $m-2$ 列の区間 $[a, b]$ と第 $m-1$ 列の区間 $[x, y]$ との関係が $a \geq x$ 及び $b \leq y$ 又は $a \leq x$ 及び $b \geq y$ である直交凸領域、又は第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、数16の値が最も大きい直交凸領域 $S^N(s, t)$ の(a)第 $m-1$ 列の区間 $[x, y]$ 及び(b)第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S^N(s, t)$ の数16の値を記憶する、という処理になる。

【0026】さらに、第4の類型のための計算は、より詳しく説明すると、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が区間 $[s, t]$ と同一である直交凸領域、第 m 列の区間 $[s, t]$ が右端列であって第 $m-1$ 列の区間 $[x, y]$ が $x \leq s$ 及び $y \geq t+1$ 又は $x \leq s-1$ 及び $y \geq t$ を満たす直交凸領域、又は第 m 列の区間 $[s, t]$ のみで構成される直交凸領域のうち、数16の値が最も大きい直交凸領域 $S^N(s, t)$ の(a)第 $m-1$ 列の区間 $[x, y]$ 及び(b)第 $m-1$ 列の区間 $[x, y]$ と第 $m-2$ 列の区間 $[a, b]$ との関係、又は第 m 列が左端列であることを示す情報を m 及び $[s, t]$ に対応して記憶手段 H^N に記憶し、当該直交凸領域 $S^N(s, t)$ の数16の値を記憶する、という処理になる。

【0027】また、最後に切り出すステップは、最初に、全ての m 及び $[s, t]$ について第1乃至第4記憶ステップを実行し、計算された数16の値のうち最大の値を有する直交凸領域 S の m 及び $[s, t]$ と、対応する記憶手段 H^N 、 H^N 、 H^N 、又は H^N とを用いて、第 $m-1$ 列の区間 $[x, y]$ 及び第 $m-1$ 列と第 $m-2$ 列との関係を読み出す(第1関係読出ステップ)。そして、第 $m-1$ 列と第 $m-2$ 列との関係を用いて、記憶手段 H^N 、 H^N 、 H^N 、及び H^N から対応する記憶手段を選択し(選択ステップ)、第 $m-1$ 列の区間 $[x, y]$ を用いて選択された記憶手段から第 $m-2$ 列の区間 $[a, b]$ 及び第 $m-2$ 列と第 $m-3$ 列との関係を読み出す(第2関係読出ステップ)。最後に選択ステップと第2関係読出ステップとを、前列との関係が「左端列であることを示す情報」となるまで繰り返す。

【0028】以下の説明を理解すれば、上述の方法を実施するような装置を作成すること、またこのような方法をコンピュータに実施させるプログラムを作成することは容易に実施できるであろう。また、上記のようなプログラムを記憶媒体や記憶デバイスに記憶することは、通常行われることである。

【0029】

【発明の実施の形態】まず、本発明の各ステップがどのように実施されるかを示す。

(1) 平面構成ステップ

先に述べたように、あるデータの2つの数値属性に2つの座標軸（x軸、y軸）をそれぞれ対応させ、これら2軸の張る平面を考える。この平面を軸ごとにN個のピクセルに分割し、平面上にN²個のピクセルを作成する。

図1に、この平面構成ステップのフローを示す。ステップ100にて処理が開始し、まずデータ集合Pからデータのランダムサンプリングを行う（ステップ110）。サンプリングされたデータをpk(xk, yk) (k = 1, 2, . . . X。xk, ykはデータの2つの数値属性の値を、Xはサンプリングされたデータ数をそれぞれ示す。)と表す。そして、xk, ykごとにソートを行い

(ステップ120)、xk, ykごとに、i・X・N (i = 1, 2, . . . N-1) 番目に小さな値を見つけ出す(ステップ130)。見つけ出された値が、各軸のピクセルの境界値となる。このようなステップを実施することにより、平面上の各列及び各行に属するデータの数はほぼ均等になる。そして、見つけ出された値を用いて、ピクセル(i, j)に入るデータpkの数u(i, j)と、その中で真偽をとる属性が真であるデータpkの数v(i, j)とをカウントする(ステップ140)。ここで、u(i, j)及びv(i, j)は、上記平面上のi行j列に存在するピクセルのデータを表すので、x軸方向にj、y軸方向にi進んだ場所にあるピクセルを表すことになる点に注意する。最後に、カウントされたu(i, j)及びv(i, j)を各ピクセルごとに記憶する(ステップ150)。このようにして、2つのN×N 30 行列、u(i, j)及びv(i, j)が生成される。

【0030】上述のようにランダムサンプリングを行うのは、通常全てのデータをソートしていると時間がかかるからである。但し、ソートしてもよいような場合もある。また、ランダムサンプリングで取り出されるデータ*

$$\begin{aligned} \Gamma_n(s, t) &= \sum_{k=s}^t g(k, n) \\ &= \begin{cases} v'(t, n) - \theta u'(t, n) & (s=0) \\ (v'(t, n) - v'(s-1, n)) - \theta (u'(t, n) - u'(s-1, n)) & (s \geq 1) \end{cases} \end{aligned}$$

【0034】さらに、

【数25】

$$U_{sum} = \sum_{全体} u$$

【数26】

$$V_{sum} = \sum_{全体} v$$

も後によく用いるので用意する。以上のような準備をすれば、以下の領域切り出しステップが高速になる。

*の数は、30Nから50Nぐらいが好ましい。また、2軸ともN個に分割する例を示したが、異なる数に分割することも可能である。典型的な例で、Nは20から1000ぐらいである。

【0031】以上述べたのは一例であって、他の方法を用いてもよい。例えば、各ピクセルの境界数値については予め定めた値を用いても良い。また、データ値に対して均等に分割することも、また対数的に分割することも可能である。

【0032】また、後の処理のため以下のような処理(図2)を行っておくと、さらに全体の処理が高速化される。すなわち、u(i, j)とv(i, j)の行数(N_v)と列数(N_u)を調べる(ステップ210)。そして、先に求めたu(i, j)とv(i, j)を用いて、新たに以下のようなu'(i, j)とv'(i, j)という行列を作成する(ステップ220)。

【数21】

$$u'(i, j) = \sum_{k=0}^i u(k, j)$$

【数22】

$$v'(i, j) = \sum_{k=0}^i v(k, j)$$

これらの計算は、全てのi = 0, 1, . . . N_v - 1, j = 0, 1, . . . N_u - 1について実施する。

【0033】このu'(i, j)とv'(i, j)は、後々数多く計算することとなる目的関数(以下、ゲインということもある)、

【数23】

$$g(i, j) = v(i, j) - \theta u(i, j)$$

の和計算を以下のように簡単化するために用意する。

【数24】

【0035】(2) 領域切り出しステップ

このステップは直交凸領域を先に作成した平面から切り出すものである。直交凸領域の例を図3に示す。直交凸領域は、(1) y軸に平行な線との交わりが必ず連続か空であって、且つ(2) x軸に平行な線との交わりが必ず連続か空な領域を言う。図3の左側の領域は、y軸に平行な、いかなる線との交わりも連続か空であり、且つx軸に平行な、いかなる線との交わりも連続か空であるので直交凸領域であると言える。一方、図3の右側の領

域は、x 軸に平行な線との交わりは必ず連続又は空であるが、y 軸に平行な線での交わりは連続でないものを含んでいる。よって、この領域は直交凸ではない。

【0036】先に示した直交凸領域の条件(1)のみを満たす領域をx単調な領域と言い、(2)のみを満たす領域はy単調な領域という。先に示した従来技術では、x単調な領域を切り出すものであったが、実際に本発明のようなデータマイニングにおいて当該従来技術を適用すると、縦に激しく揺れる入り組んだ領域を切り出すことが多く、人間には理解し難い形状となってしまう。また、任意の形状で切り出そうとすると、その問題はNP困難となってしまう。そこで、本発明では、直交凸領域で切り出すこととする。

【0037】直交凸領域を切り出す際には、パラメータ値 θ (0以上1以下の実数)を含む数23で表されるゲインを領域全体で最大にするような直交凸領域を切り出す。ここで、パラメータ θ の説明をしておく。図4に示すような、横軸が切り出される領域Sに含まれるデータ数 $U(S)$ 、縦軸が切り出される領域Sに含まれ且つ真偽をとる属性が真であるデータの数 $V(S)$ であるような平面を考える。データ数と真偽をとる属性が真であるデータの数の組み合わせは多数存在するので、この平面には多数の点が存在することになるが、この点のうち、凸包を構成する点を特に用いる。すなわち、この凸包を構成する点をつなぐことにより曲線を構成し、この曲線に対し傾き θ を有する直線を上から下ろして行き、最初にこの曲線と接する点を求め、この時の領域を出力する。凸包上の点は図4では黒丸で表されている。以下、凸包上の点をフォーカス・イメージという。また、直線を下ろしていくような方法をハンドプロープという。このように、本発明では傾き θ をパラメータとして入力するような方法を用いる。

【0038】このように凸包上の点のみ取り扱うのは、コンフィデンス最大化ルール、サポート最大化ルールは、凸包上に必ず存在するわけではないが、近似解としては十分な点を出力することができ、また最適化エントロピー・ルール及び最適化インタクラスバリエーション・ルールについては、この凸包上に必ず存在するからである。もし、コンフィデンス最大化ルール及びサポート最大化ルールの厳密解を解くとすると、実用的な時間には計算が終了しないので、近似解であっても十分に有効な結果を出力できる。

【0039】上記のように傾き θ の直線を下ろしていくということは、直線 $y = \theta x + Q$ のY切片である Q を減少させることであり、言い換えれば、 $Q = V(S) - \theta U(S)$ を最大にする $U(S)$ をX座標に有する点を求める問題となる。よって、

【数27】

10

20

30

40

50

20

$$\begin{aligned} \max Q &= \max \left\{ \sum_{(i,j) \in S} v(i,j) - \theta \sum_{(i,j) \in S} u(i,j) \right\} \\ &= \max \sum_{(i,j) \in S} g(i,j) \end{aligned}$$

と変形される。

【0040】では、この数27をどのように解くかを考える。最初に、直交凸領域の性質を領域内部のピクセル同士の関係によって表現する。Sをピクセル平面内の直交凸領域とする。 m_l 、 m_r をそれぞれSの左端、右端の列番号とする。Sの第m列($m_l \leq m \leq m_r$)の下端及び上端のピクセル番号をそれぞれ $s(m)$ 、 $t(m)$ とする。これらの位置関係は図5を参照するとよく分かる。第m列の区間 $[s(m), t(m)]$ の変化傾向を第m-1列の区間との比較によって次のように定義する。

(a) W-Type: 広がり型 (図6左上)

$s(m-1) \geq s(m)$ 、 $t(m-1) \leq t(m)$ の場合

(b) U-Type: 上昇型 (図6右上)

$s(m-1) \leq s(m)$ 、 $t(m-1) \leq t(m)$ の場合

(c) D-Type: 下降型 (図6左下)

$s(m-1) \geq s(m)$ 、 $t(m-1) \geq t(m)$ の場合

(d) N-Type: 狭まり型 (図6右下)

$s(m-1) \leq s(m)$ 、 $t(m-1) \geq t(m)$ の場合

【0041】 $m=m_l$ の列は全変化傾向に属し、上の不等式の等号が成り立つ場合、その列は複数の変化傾向に同時に属する。上の定義から直交凸領域内のどの列の区間も上の4種類のタイプのいずれかに属する。

【0042】また、直交凸領域の性質から、ある変化傾向の列の左隣の列の変化傾向は次の条件を満たす。すなわち、

(1) W-Typeの左隣の列はW-Typeである。

(2) U-Typeの左隣の列はW-Type又はU-Typeである。

(3) D-Typeの左隣の列はW-Type又はD-Typeである。

(4) N-Typeの左隣の列はW-Type、又はU-Type、又はD-Type、又はN-Typeである。

【0043】このような条件を満たす領域は逆に言うとき直交凸領域であると言える。これらの条件は、図7に状態遷移図として示されている。図中のW、U、D、Nは、それぞれW-Type、U-Type、D-Type、N-Typeであり、矢印を1つたどるごとに1つ右隣の列の状態に遷移する。

【0044】全ての直交凸領域は領域の右端の列の変化傾向によって先に示した4つの種類に分類できる。ここ

で、4つの種類を総称してXタイプ($X \in \{W, U, D, N\}$)と呼ぶ。列の区間のタイプと同様に、領域のタイプも複数のタイプに同時に属する場合もある。

【0045】また、右端が第m列の区間 $[s, t]$ であ*

$$f_m^X(s, t) = \max \{f_m^W(s, t), f_m^U(s, t), f_m^D(s, t), f_m^N(s, t)\}$$

である。この $f_m^X(s, t)$ を $m=0, \dots, N_x-1$ ($\forall (s \leq t)$)について求めて、それらの中で最大のものを選びだせば、それが先の平面内の全ての直交凸領域のゲインの最大値となる。

【0046】この最大値を求めるために、 $m=0, \dots, N_x-1$ に対して順番に、 $f_m^X(s, t)$ ($\forall (s \leq t)$)を全て計算するという方針をとる。

【0047】次に $m=0$ 、すなわち第1列の $f_0^X(s, t)$ を計算する。この場合、全てのタイプで同一である。これは、

【数29】

$$\begin{aligned} f_0^X(s, t) &= \Gamma_0(s, t) \\ &\quad \forall X \in \{W, U, D, N\} \\ &\quad \forall (s \leq t) \end{aligned}$$

※

$$f_m^W(s, t) = \max \begin{cases} \Gamma_m(s, t) & (1) \\ f_{m-1}^W(s, t) + \Gamma_m(s, t) & (2) \\ f_m^W(s, t-1) + g(t, m) & (s < t) & (3) \\ f_m^W(s+1, t) + g(s, m) & (s < t) & (4) \end{cases}$$

ここで、 \max を求める時に、 $s=t$ の場合は数30の(1)(2)式だけで比較をし、大きい方の値を用いる。その他の場合は(1)乃至(4)のすべてから最も大きい値を用いる。

【0050】数30の(1)式は、第m列の区間 $[s, t]$ だけからなる領域(幅1の縦長の長方形)のゲインを表す。また(2)式は、第m-1列がW-Typeでその区間が $[s, t]$ 、且つ第m列も区間 $[s, t]$ で右端となっている領域のうち最大のゲインを表す。これは、図8(a)に表したような場合を示す。なお、直交凸領域の性質から第m列がW-Typeであれば第m-1列がW-Typeであることは決まる。

【0051】また(3)式は、第m-1列がW-Typeであって、その区間 $[s(m-1), t(m-1)]$ が、 $s(m-1) \geq s, t(m-1) \leq t-1$ を満たし、第m列の区間 $[s, t]$ で右端という領域のうち最大のゲインを示す。これは、図8(b)のような形状を意味する。第m-1列の上端は、 $t-1$ 以下であり、下端は s 以上である。(4)式は、第m-1列がW-Typeであって、その区間 $[s(m-1), t(m-1)]$ が $s(m-1) \geq s+1, t(m-1) \leq t$ を満たし、第m列の区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは、図8(c)のような形状を意味する。第m-1列の上端は t 以下であり、下端は $s+1$ 以上である。(2)乃至(4)式は、W-Typeの左列はW-Typeしかあり得ないということが考慮されている。

※るXタイプの直交凸領域のゲインの最大値を $f_m^X(s, t)$ と表す。そして、4つのタイプ領域のゲインのうち最も大きいものを、 $f_m^X(s, t)$ と表す。すなわち、

【数28】

※で求められる。

【0048】そして、 $f_{m-1}^X(s, t)$ ($\forall X \in \{W, U, D, N\}, (\forall (s \leq t))$)を求める。以下は、各タイプごとに説明する。

【0049】(a) 広がり型(W-Type)の場合第m列の区間 $[s, t]$ を最右端とする直交凸領域であって第m列がW-Typeである領域のゲインの最大値 $f_m^W(s, t)$ は、以下の式により求められる。

【数30】

1)】が $s(m-1) \geq s+1, t(m-1) \leq t$ を満たし、第m列の区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは、図8(c)のような形状を意味する。第m-1列の上端は t 以下であり、下端は $s+1$ 以上である。(2)乃至(4)式は、W-Typeの左列はW-Typeしかあり得ないということが考慮されている。

【0052】以上の $f_m^X(s, t)$ の計算を一系列中の全ての区間 $[s, t]$ に対して行う。この計算は図9のアルゴリズムに従う。以上のように、第m列が領域の右端で第m-1列からの変化傾向がW-Typeという領域のうち最大のゲインが得られる。

【0053】(b) 上昇型(U-Type)の場合最初に以下の式の値を求めておく。

【数31】

$$\beta_{m-1}^W(s, t) = \{i \mid \max_{i \leq s} f_{m-1}^W(i, t)\}$$

【数32】

$$\beta_{m-1}^U(s, t) = \{i \mid \max_{i \leq s} f_{m-1}^U(i, t)\}$$

* $t]$ を最右端とする直交凸領域であって第 m 列が U - T y p e である領域のゲインの最大値 $f_m^U(s, t)$ は、以下の式により求められる。

【数 3 3】

これは、図 1 0 のようなアルゴリズムにて実行される。

【0 0 5 4】以上の計算を用いて、第 m 列の区間 $[s, *$

$$f_m^U(s, t) = \max \begin{cases} \Gamma_m(s, t) & (1) \\ f_{m-1}^W(\beta_{m-1}^W(s, t), t) + \Gamma_m(s, t) & (2) \\ f_{m-1}^U(\beta_{m-1}^U(s, t), t) + \Gamma_m(s, t) & (3) \\ f_m^U(s, t-1) + g(t, m) & (s < t) \quad (4) \end{cases}$$

(1) 乃至 (3) 式は、 $s = t$ の場合に比較に用いられ、その際 (4) 式は用いられない。

【0 0 5 5】数 3 3 の (1) 式は、第 m 列の区間 $[s, t]$ のみからなる領域（幅 1 の縦長の長方形）のゲインを表す。また、(2) 式は、第 $m-1$ 列が W - T y p e であって、その区間 $[s(m-1), t(m-1)]$ が、 $s(m-1) \leq s, t(m-1) = t$ を満たし、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは、図 1 1 (a) に示した形状の場合であって、第 $m-1$ 列の下端の上限は s である。

【0 0 5 6】(3) 式は、第 $m-1$ 列が U - T y p e であって、その区間 $[s(m-1), t(m-1)]$ が、 $s(m-1) \leq s, t(m-1) = t$ を満たし、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは、図 1 1 (b) に示した形状の場合であって、第 $m-1$ 列の下端の上限は s である。(4) 式は、第 $m-1$ 列が W - T y p e 又は U - T y p e であって、その区間 $[s(m-1), t(m-1)]$ が、 $s(m-1) \leq s, s \leq t(m-1) \leq t-1$ を満たし、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは、図 1 1 (c) に示した形状の場合であって、第 $m-1$ 列の下端の上限は s であり、また上端の範囲は s 以上 $t-1$ 以下である。(2) 乃至 (4) 式は、 U - T y p e の左隣列は W - T y p e 又は U - T y p e しかあり得ないということが考慮されてい※

※る。

【0 0 5 7】以上の $f_m^U(s, t)$ の計算を一系列中の全ての区間 $[s, t]$ に対して行う。この計算は、図 1 2 に示すアルゴリズムに従う。このように、第 m 列が領域の右端でその変化傾向が U - T y p e という領域のうち最大のゲインが得られる。

【0 0 5 8】(c) 下降型 (D - T y p e) の場合最初に以下の式の値を計算しておく。

【数 3 4】

$$\tau_{m-1}^W(s, t) = \{i \mid \max_{i \geq t} f_{m-1}^W(s, i)\}$$

【数 3 5】

$$\tau_{m-1}^D(s, t) = \{i \mid \max_{i \geq t} f_{m-1}^D(s, i)\}$$

これらの計算は、図 1 3 に示されたアルゴリズムにより実行される。(1) 乃至 (3) 式は、 $s = t$ の場合に比較に用いられ、その際 (4) 式は用いられない。

【0 0 5 9】以上の計算を用いて、第 m 列の区間 $[s, t]$ を最右端とする直交凸領域であって第 m 列が D - T y p e である領域のゲインの最大値 $f_m^D(s, t)$ は、以下の式により求められる。

【数 3 6】

$$f_m^D(s, t) = \max \begin{cases} \Gamma_m(s, t) & (1) \\ f_{m-1}^W(s, \tau_{m-1}^W(s, t)) + \Gamma_m(s, t) & (2) \\ f_{m-1}^D(s, \tau_{m-1}^D(s, t)) + \Gamma_m(s, t) & (3) \\ f_m^D(s+1, t) + g(s, m) & (s < t) \quad (4) \end{cases}$$

【0 0 6 0】数 3 6 の (1) 式は、第 m 列の区間 $[s, t]$ のみからなる領域（幅 1 の縦長の長方形）のゲインを表す。また、(2) 式は、第 $m-1$ 列が W - T y p e

であって、その区間 $[s(m-1), t(m-1)]$ が、 $s(m-1) = s, t(m-1) \geq t$ を満たし、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲ

インを表す。これは、図14(a)に示した形状の場合であって、第 $m-1$ 列の上端の下限は t である。

【0061】(3)式は、第 $m-1$ 列がD-Typeであって、その区間 $[s(m-1), t(m-1)]$ が、 $s(m-1) = s, t(m-1) \geq t$ を満たし、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは図14(b)に示した形状の場合であって、第 $m-1$ 列の上端の下限は t である。(4)式は、第 $m-1$ 列がW-Type又はD-Typeであって、その区間 $[s(m-1), t(m-1)]$ が、 $s+1 \leq s(m-1) \leq t, t(m-1) \geq t$ を満たし、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは図14(c)に示した形状の場合であって、第 $m-1$ 列の上端の下限は t であって、下端の範囲*

$$f_m^N(s, t) = \max \left\{ \begin{array}{ll} \Gamma_m(s, t) & (1) \\ f_{m-1}^W(s, t) + \Gamma_m(s, t) & (2) \\ f_{m-1}^U(s, t) + \Gamma_m(s, t) & (3) \\ f_{m-1}^D(s, t) + \Gamma_m(s, t) & (4) \\ f_m^N(s, t+1) + \Gamma_m(s, t) & (5) \\ f_m^N(s, t+1) - g(t+1, m) & (t < N_y - 1) \quad (6) \\ f_m^N(s-1, t) - g(s-1, m) & (s > 0) \quad (7) \end{array} \right.$$

ここで、 \max を求める時、各式は式の後ろの条件を満たす場合にのみ用いられる。すなわち、(6)式は $t > N_y - 1$ を満たす時のみ比較され、(7)式は $s > 0$ を満たす場合にのみ比較に用いられる。

【0064】数37の(1)式は、第 m 列の区間 $[s, t]$ のみからなる領域(幅1の縦長の長方形)のゲインを表す。(2)式は、第 $m-1$ 列がW-Typeであって、その区間が $[s, t]$ であり、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは図16(a)に示した形状の場合である。(3)式は、第 $m-1$ 列の区間 $[s, t]$ がU-Typeであって、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは図16(b)に示した形状の場合である。(4)式は、第 $m-1$ 列の区間 $[s, t]$ がD-Typeであって、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは図16(c)に示した形状の場合である。(5)式は、第 $m-1$ 列の区間 $[s, t]$ がN-Typeであって、その区間 $[s, t]$ であり、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは図16(d)に示した形状の場合である。

*は $s+1$ 以上 t 以下である。(2)乃至(4)式は、D-Typeの左隣列はW-Type又はD-Typeしかあり得ないということが考慮されている。

【0062】以上の $f_m^N(s, t)$ の計算を一系列中の全ての区間 $[s, t]$ に対して行う。この計算は図15のアルゴリズムに従う。このようにして、第 m 列が領域の右端でその変化傾向がD-Typeという領域のうち最大のゲインが得られる。

【0063】(d)狭まり型(N-Type)の場合第 m 列の区間 $[s, t]$ を最右端とする直交凸領域であって第 m 列がN-Typeである領域のゲインの最大値 $f_m^N(s, t)$ は、以下の式により求められる。

【数37】

【0065】(6)式は、第 $m-1$ 列がW-Type、U-Type、D-Type又はN-Typeであって、その区間 $[s(m-1), t(m-1)]$ が、 $s(m-1) \leq s, t(m-1) \geq t+1$ を満たし、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは図16(e)に示した形状であって、第 $m-1$ 列の上端の下限は $t+1$ であり、下端の上限は s である。(7)式は、第 $m-1$ 列がW-Type、U-Type、D-Type又はN-Typeであって、その区間 $[s(m-1), t(m-1)]$ が、 $s(m-1) \leq s-1, t(m-1) \geq t$ を満たし、第 m 列は区間 $[s, t]$ で右端という領域のうち最大のゲインを表す。これは図16(f)に示した形状であって、第 $m-1$ 列の上端の下限は t であり、下端の上限は $s-1$ である。

【0066】以上の $f_m^N(s, t)$ の計算を一系列中の全ての区間 $[s, t]$ に対して行う。この計算は次のアルゴリズムに図17に従う。このようにして、第 m 列の区間 $[s, t]$ が領域の右端でその変化傾向がN-Typeという領域のうち最大のゲインが得られる。

【0067】上述の(a)乃至(d)の計算にて各列の

各 $[s, t]$ を右端とする領域の最大のゲインを計算することができる訳であるが、それと同時に“領域”自体も同時に記録しておく必要がある。これは、後の出力ステップでは、この求められた直交凸領域内に含まれるデータを取り出すからである。

【0068】ここで、同じ最大値の領域が複数存在する場合には、それらのうち先に見つけた方を解として取り扱う。また、領域は縦方向の区間が横に並んだものとして

$$[s(m_1), t(m_1)], \dots, [s(m_r), t(m_r)] \quad 10$$

のように表現する。1列は N_y 行あるので、区間 $[s, t]$ ($s \leq t$) の総数は $N_y(N_y + 1) / 2$ 個ある。各区間 $[s, t]$ は1つの整数 p に一対一対応させる。

【0069】例えば、次の関数 $h(s, t)$ によって区間 $[s, t]$ を1つの整数 $p = h(s, t)$ に対応させることができる。すなわち、

【数38】

$$h(s, t) = s + \frac{t(t+1)}{2}$$

である。

【0070】逆に、区間を表す整数 p ($0 \leq p \leq N_y(N_y + 1) / 2$) から区間 $[s, t]$ は、次のように求められる。

【数39】

$$t = \left\lfloor \frac{-1 + \sqrt{1 + 8p}}{2} \right\rfloor$$

【数40】

$$s = p - \frac{t(t+1)}{2}$$

但し、このような計算はどの領域が最大のゲインを有するのかを決定する際には用いることはない。以後、

$[s, t]$ は1つの整数と同一視して取り扱う。また、上記数38は一例にすぎず、他の関数を用いても問題ない。

【0071】次に領域を記憶しておく配列を用意する。これは、 $N_x \times N_y(N_y + 1) / 2$ の整数型2次元配列であり、W, U, D, N-Typeのそれぞれに対して1つ用意する。この要素 $H^x(m, [s, t])$ ($0 \leq m \leq N_x - 1$, $0 \leq [s, t] \leq N_y(N_y + 1) / 2$, $X \in \{W, U, D, N\}$) と表すこととする。

【0072】この要素 $H^x(m, [s, t])$ には、ゲインが $f_{\pm}^x(s, t)$ の領域の第 $m-1$ 列の区間 $[x, y]$ と、第 $m-1$ 列の第 $m-2$ 列からの変化傾向 Y を表す数値を記憶する。以下、 $H^x(m, [s, t]) = Y : [x, y]$ と表す。例えば、この $H^x(m, [s, t])$ を整数型32ビットで表現し、 Y の部分を上位2ビット、残りの下位ビットを $[x, y]$ を表すのに用いる(図18参照)。

【0073】但し、第 m 列が領域の左端列である場合

に、第 $m-1$ 列にはつながないことを表すために、この下位ビットには領域の左端を表す値を入れる。例えば、先ほどの数40で区間を表現する例では、この下位ビットに $N_y(N_y + 1) / 2$ 以上の値を入れるか又は29ビット目を領域の左端を表すフラグにすればよい。

【0074】では、最終的に最大のゲインを有する直交凸領域を求める処理を図19を用いて説明する。ステップ1600で開始された処理は、最初に $m=0$ として、 m を初期化する(ステップ1610)。次に、 $m=N_x$ であるか判断する(ステップ1620)。これは、 m が N_x に達して、全ての列について以下の計算が終了したかを判断するものである。もし、全ての列 m について計算が終了していなければ、全ての $[s, t]$ について、 $H^{\pm}(m, [s, t])$ と $f_{\pm}^{\pm}(s, t)$ 、 $H^{\pm}(m, [s, t])$ と $f_{\pm}^{\pm}(s, t)$ 、 $H^{\pm}(m, [s, t])$ と $f_{\pm}^{\pm}(s, t)$ 、 $H^{\pm}(m, [s, t])$ と $f_{\pm}^{\pm}(s, t)$ を計算し、その結果を記憶する。この計算の順番は任意である。そして、各計算中それまでに計算されたゲインの最大値より大きい値が計算されたならば、その値及びその $m, [s, t], X$ を記憶しておく(ステップ1630)。

【0075】ここで、 $H^{\pm}(m, [s, t])$ と $f_{\pm}^{\pm}(s, t)$ の計算は、先に示した数30の計算を実施すればよい。よって、 $H^{\pm}(m, [s, t])$ は、数30の(1)式が最大であれば領域の左端を表す値、(2)式が最大であれば $W : [s, t]$ 、(3)式が最大であれば $H^{\pm}(m, [s, t-1])$ 、(4)式が最大であれば $H^{\pm}(m, [s+1, t])$ となる。以上のように、 $H^{\pm}(m, [s, t])$ のみを考えれば、前列である第 $m-1$ 列は、必ず $W-Type$ であるから、第 $m-1$ 列の第 $m-2$ 列からの変化傾向は記憶する必要ない。

【0076】また、 $H^{\pm}(m, [s, t])$ と $f_{\pm}^{\pm}(s, t)$ の計算は、先に示した数33の計算を実施すればよい。よって、 $H^{\pm}(m, [s, t])$ には、数33の(1)式が最大であれば領域の左端を表す値、(2)式が最大であれば $W : [\beta_{m-1}^{\pm}(s, t), t]$ 、(3)式が最大であれば $U : [\beta_{m-1}^{\pm}(s, t), t]$ 、(4)式が最大であれば $H^{\pm}(m, [s, t-1])$ が記憶される。

【0077】 $H^{\pm}(m, [s, t])$ と $f_{\pm}^{\pm}(s, t)$ の計算は、先に示した数36の計算を実施すればよい。よって、 $H^{\pm}(m, [s, t])$ には、数36の(1)式が最大であれば領域の左端を表す値、(2)式が最大であれば $W : [s, \tau_{m-1}^{\pm}(s, t)]$ 、(3)式が最大であれば $D : [s, \tau_{m-1}^{\pm}(s, t)]$ 、(4)式が最大であれば $H^{\pm}(m, [s+1, t])$ が記憶される。

【0078】最後に、 $H^{\pm}(m, [s, t])$ と $f_{\pm}^{\pm}(s, t)$ の計算は、先に示した数37の計算を実施すればよい。よって、 $H^{\pm}(m, [s, t])$ は、数37の(1)式が最大であれば領域の左端を表す値、(2)式が最大であれば $W : [s, t]$ 、(3)式が最大であれば $U : [s, t]$ 、(4)式が最大であれば $D : [s, t]$ 、(5)式が最大であれば $N : [s, t]$ 、(6)式が最大であれば H

20

30

40

50

$H^N(m, [s, t+1])$ 、(7) 式が最大であれば $H^N(m, [s-1, t])$ となる。

【0079】ここまでで分かるように、すべての $f^X(s, t)$ を記憶しておく必要はない。第 m 列の計算を実施している時には、その第 m 列と第 $m-1$ 列の計算結果のみを用いる。よって、 W 、 U 、 D 、 $N-Type$ ごとに 2 列分の記憶容量があればよい。但し、余裕があれば全て記憶しておいてもよい。

【0080】図 19 のステップ 1630 を終了すると、 m を 1 インクリメントして (ステップ 1640)、ステップ 1620 に戻る。そして、この処理を全ての列について実施する。もし、全ての列について実施されたならば、全ての列に関して最大のゲイン値を有していた領域に関する m 、 $[s, t]$ 、 X から、 $H^X(m, [s, t])$ を参照し、その値 $Y: [x, y]$ を取り出す (ステップ 1650)。ここまでの処理で、最右端列である第 m 列と、その列の区間 $[s, t]$ 、第 $m-1$ 列とその区間 $[x, y]$ が分かる。

【0081】次に、第 $m-1$ 列で領域は左端となる場合もあるので、 $[x, y]$ が左端を表す値であるか判断される (ステップ 1660)。左端であれば、ここで処理は終了する (ステップ 1680)。左端でなければ、 Y を X として、 $[x, y]$ を $[s, t]$ として、 $m-1$ を m とし (ステップ 1670)、ステップ 1650 に戻る。このように、 $[x, y]$ が左端を表す値となるまで、この処理を繰り返せば、最大のゲイン値を有する直交凸領域の各列の区間を得ることができる。

【0082】(3) 出力ステップ

以上のように求まった直交凸領域 S は、前記平面のどの部分を占めているかは、先のステップによりわかっているので、その領域 S に属するデータを取り出すことになる。通常各データは、真偽をとる属性及び数値属性のみならず、他の属性も有しているから、例えばダイレクトメールを送るのであれば、住所氏名といった属性を取り出すようになる。ここまでくると、取り出すべきデータは特定されているから、通常のデータベースの検索に過ぎないので、これ以上詳しく述べない。当然、一旦直交凸領域をその外形がよくわかるようにして、ユーザに提示するようにしてもよい。

【0083】以上のような各ステップを実施すれば、ある条件 θ に対する、データ間結合ルール of 1 つを求めることができる。しかし、この条件 θ をどのように設定するかということは、1 つの問題である。通常、ある条件 θ 1 つでは、問題の解決にならない場合が多い。以上の各ステップ、特に (2) 領域切り出しステップをエンジンとして用い、どのように先に述べた 4 つの一般的なルール及び他のルール等を導き出すかを以下に示す。

【0084】A. ある区間に存在する直交凸領域を求める場合

まず、幾つかの θ に対応するフォーカス・イメージ S を

連続的に示し、動画を作成することにより、切り出される領域の大きさ及び形状をユーザの判断により決定させる場合を考える。

【0085】この処理を図 20 に示す。ステップ 800 にて開始された処理は、まず $\theta 1$ を入力することにより、上述したプロセスにてフォーカス・イメージ $S 1$ を見つけ出す (ステップ 810)。また、ユーザに $\theta 2$ を入力させ、同様にフォーカス・イメージ $S 2$ を見つけ出す (ステップ 820)。このようにして 2 つのフォーカス・イメージが求まると、それぞれに含まれるデータ数 $U(S 1)$ 、 $U(S 2)$ 及び真偽をとる属性が真であるデータの数 $V(S 1)$ 、 $V(S 2)$ とを用いて、その中間にある、新たな傾き $\theta 3$ を計算する (ステップ 830)。

【0086】このように新たな $\theta 3$ が求まれば、さらにこの $\theta 3$ に対応するフォーカス・イメージ $S 3$ を求めることができる (ステップ 840)。ここで、計算された $S 3$ が既に求まっていれば、区間 $(\theta 1, \theta 2)$ にはこれ以上のフォーカス・イメージは凸包上 (図 4) には存在しない。よって、処理が終了する (ステップ 880)。

しかし、発見済みでなければ、 $\theta 2$ の代わりに $\theta 3$ を用いて、ステップ 830 以降を実行する (ステップ 860)。すなわち、区間 $(\theta 1, \theta 3)$ の間にあるフォーカス・イメージを見つけ出す。この場合、次々に中間の値を計算していくようにすることも可能である。また、ある程度の個数フォーカス・イメージが求まったところで計算を取り止めることもできる。さらに、もう 1 つ残った区間 $(\theta 3, \theta 2)$ についてフォーカス・イメージを計算するために、 $\theta 3$ 、 $\theta 2$ についてステップ 830 以降を実行する (ステップ 870)。この場合も、この区間内に存在しているフォーカス・イメージを全て見つけ出すようにしてもよいし、所定の個数見つけ出したところで処理を終了してもよい。

【0087】このようにして、1 つ又は複数のフォーカス・イメージを見つけ出すことができた。このように求まった複数のフォーカス・イメージを連続してユーザに提示するようなことも可能である。

【0088】B. コンフィデンス最大化ルールの場合 (図 21 及び図 22)

この場合には、ルールの定義より最小限度のサポート $minsup$ (全体のデータ数に対する領域に包含されるデータ数の割合) を入力する (ステップ 910)。ここで、 $Umin = Usum \times minsup$ を計算しておく。ここで図 4 を見てみると、最小限度サポートと記された縦の点線がこの値に対応する。まず、 $\theta = 1$ でフォーカス・イメージ $S 1$ を求める (ステップ 920)。そして、この $S 1$ に含まれるデータ数 $U(S 1)$ が、 $U(S 1) \geq Umin$ を満たすかどうか判断する (ステップ 930)。もし成立するならば、 $S 1$ を解として決定し (ステップ 950)、処理を終了する (ステップ 990)。成り立たない場合、フォーカス・イメージ $S 2$ を平面全体を表すイメージとす

10

20

30

40

50

る。すなわち、 $U(S2) = U_{sum}$ 、 $V(S2) = V_{sum}$ と代入する(ステップ940)。そして、 XX を介して図22に移行する。

【0089】図22では、 XX から始まり、新たな条件 θ を求め、この θ に対するフォーカス・イメージ S を計算する(ステップ1400)。この θ は

$$\theta = (V(S2) - V(S1)) / (U(S2) - U(S1))$$

にて計算される。そして、 $S1 = S$ 又は $S2 = S$ であるならば、($S1$, $S2$)の間にはこれ以上フォーカス・イメージは存在しないので、コンフィデンスの高い $S2$ が最良解として出力され、処理を終了する(ステップ1410)。また、 $U(S) \equiv U_{min}$ であるならば、 S を出力し、処理を終了する。

【0090】ところが、 $U(S) < U_{min}$ であると(ステップ1420)、まだ処理が必要なので、 $S1 = S$ として(ステップ1440)、ステップ1400に戻る。同様に、 $U(S) > U_{min}$ であるならば、 $S2 = S$ として(ステップ1430)、ステップ1400に戻る。

【0091】これを繰り返すことにより解が見つけれらる。図4を参照すると、先に説明した最小限度のサポートの右側、濃く塗られた部分に解の存在する範囲がある。そして、この図4の場合には、凸包の内部の白丸の点が厳密解となるが、本発明ではハンド・プロープにて得られた近似解が出力される。見つけられた解は、ユーザに提示されるようにしてもよいし、そのフォーカス・イメージに属するデータの必要な属性を出力するようにしてもよい。

【0092】C. サポート最大化ルールの場合(図23、図24)

この場合、ルールの定義より、最小限度のコンフィデンス $minconf$ (直交凸領域に包含されるデータ数に対する真偽をとる属性が真である割合)を入力する(ステップ1110)。図4の場合、最小限度のコンフィデンスと示され、原点から引かれた点線がこれに該当する。まず、フォーカス・イメージ $S2$ を平面全体を表すイメージとする。すなわち、 $U(S2) = U_{sum}$ 、 $V(S2) = V_{sum}$ と代入する(ステップ1120)。そして、 $minconf \leq V(S2) / U(S2)$ であるかを判断する(ステップ1130)。もしこの条件が成立するならば、 $S2$ を解として決定し(ステップ1160)、処理を終了する(ステップ1190)。条件が成立しないならば、 $\theta = 1$ でフォーカス・イメージ $S1$ を求める(ステップ1140)。そして、 $minconf > V(S1) / U(S1)$ が成り立つかどうか判断する(ステップ1150)。もし成り立つならば、解は存在せず、処理を終了する。成り立たないならば、 Y を介して図24へ移行する。

【0093】図24では、 Y から処理が開始され、 $\theta = (V(S2) - V(S1)) / (U(S2) - U(S1))$ としてフォーカス・イメージ S を求める(ステップ12

00)。この求められたフォーカス・イメージ S に対し、(1) $minconf \equiv V(S) / U(S)$ が成立する場合には、この S を出力して処理を終了する(ステップ1210)。また、 $S1 = S$ 若しくは $S2 = S$ である場合には、これ以上 $S1$ と $S2$ の間には解は無いので、 $S1$ を最良解として出力し、処理を終了する(ステップ1210)。これに対し、 $minconf < V(S) / U(S)$ である場合には(ステップ1220)、 $S1 = S$ としてステップ1200に戻る(ステップ1230)。また、 $minconf > V(S) / U(S)$ である場合には、 $S2 = S$ としてステップ1200に戻る(ステップ1240)。

【0094】以上のようにして、サポート最大化ルールが求められる。もう一度図4に戻ると、先に説明した最小限度のコンフィデンスとして示した点線より上の濃く塗られた範囲に解が存在する。そして、この例では凸包内の白丸の点が厳密解であるが、このように凸包内部の点は見つけ出すのに膨大な計算量を必要とするので、凸包上の点でサポートを最大にする近似解を出力するようにしている。先に述べたように、見出された近似解又は厳密解は、ユーザに提示してもよいし、フォーカス・イメージ内に含まれるデータの必要な属性値を出力するようにしてもよい。

【0095】D. 最適化エントロピ・ルールの場合
最適化エントロピ・ルールとは、領域の内部と外部との分割を考えた時、分割前の情報量と比較した分割後の情報量の増分を最大化するルールである。よって、切り出された領域と平面全体のエントロピのゲイン(以下の式)が最大となる領域を発見すればよい。

【数41】

$$f(x, y) = -\frac{y}{a} \log \frac{y}{a} - \frac{x-y}{a} \log \frac{x-y}{a} \\ - \frac{b-y}{a-x} \log \frac{b-y}{a-x} \\ - \frac{a-b-x+y}{a-x} \log \frac{a-b-x+y}{a-x}$$

この x は $U(S)$ 、 y は $V(S)$ 、 a は U_{sum} 、 b は V_{sum} である。このような条件においても、解は凸包上に存在することが分かったので、上述のステップを用いることができる。よって、 θ を変化させ、数41を最大化するフォーカス・イメージを求めればよい。

【0096】E. 最適化インタクラスバリエンス・ルールの場合

先に述べたように最適化インタクラスバリエンス・ルールとは、領域内外の分割を考えた時、内外の「標準化された真偽の割合の平均からのずれ」の二乗和を最大化するルールである。よって、切り出された領域と平面全体のインタクラスバリエンス(以下の式)が最大となる領域を発見すればよい。

【数42】

$$f(x, y) = x \left(\frac{y}{x} - \frac{b}{a} \right)^2 + (a+x) \left(\frac{b-a}{a-x} - \frac{b}{a} \right)^2$$

x, y, a, b は上述したものと同一である。このような条件においても、解は凸包上に存在することが分かったので、上述のステップを用いることができる。よって、 θ を変化させ、数42を最大化するフォーカス・イメージを求めればよい。

【0097】F. その他

以上述べたように、 $U(S)$ と $V(S)$ 上の凸包上の点に存在する又は存在すると近似できる場合には、上述したステップを用いれば高速にルールに該当する領域を導き出すことができる。

【0098】G. 二次的なルールの抽出

上述のプロセスを用いて1つのルールを見出した後に、二次的なルールを見つけ出すことができる。すなわち、切り出した1のフォーカス・イメージに属する $v(i, j)$ を除去し、 $v(i, j) \neq u(i, j) = Vsum / Usum$ となるように、 $v(i, j)$ を変更し、それから新たに領域切り出しステップを行うのである。

【0099】以上、本発明における処理のプロセスを説明した。このような処理プロセスは、コンピュータ・プログラムによって実現し、実行するようにしてもよい。例えば、図25のような通常のコンピュータ・システムにおいて実行できるようなプログラムにすることもできる。処理プログラムは、HDD1050に格納され、実行時にはメインメモリ1020にロードされ、CPU1010によって処理される。また、HDD1050はデータベースをも含んでおり、処理プログラムはそのデータベースに対するアクセスを行う。最初の平面やフォーカス・イメージは、表示装置1060によってユーザに提示される。ユーザは、入力装置1070にてフォーカス・イメージの選択や、データ出力の命令を入力する。このような入力装置には、キーボードやマウス、ポインティング・デバイスやディジタイザを含む。さらに、出力結果を補助記憶装置であるFDD1030のフロッピー・ディスクに記憶したり、また新たなデータをFDD1030から入力することもできる。さらに、CD-ROMドライブ1040を用いて、データを入力することもできる。

【0100】さらに、本発明の処理プロセスを実現したコンピュータ・プログラムは、フロッピー・ディスクやCD-ROMといった記憶媒体に記憶して、持ち運ぶことができる。この場合、通常のデータベース検索プログラムのデータ取り出し部分や、表示装置1060に表示するだけの処理を行うプログラムは、すでにHDD1050に記憶されている場合もある。よって、それ以外の部分が、上記のような記憶媒体にて流通することは通常行われる事項である。また、図示されていない通信装置がバス1080に接続されており、遠隔地にあるデータ

ベースを用いて処理したり、処理結果を遠隔地に送信するようにしてもよい。

【0101】また、本発明の処理を実施する特別の装置を設けてもよい。例えば、図26のような装置が考えられる。平面構成装置1310は、データベース1300及び切出装置1320に接続されており、制御装置1340からの命令を受付ける。また、切出装置1320は、出力デバイス1330及び表示装置1350に接続されており、制御装置1340からの命令を受付ける。また、切出装置1130はデータベースにも接続を有している。制御装置1340は、入力デバイス1360に接続され、入力デバイス1360により指示された処理の種類により平面構成装置1310及び切出装置1320を制御する。

【0102】この装置の簡単な動作を説明する。平面構成装置1310は、先に説明した平面構成ステップを実行する部分である。このように平面構成装置1110は、データベースに記憶されたデータを用いて先に示した平面を構成し、切出装置1320に出力する。切出装置1320は、制御装置1340からの命令に従って、切り出しのためのパラメータである θ をセットする。セットされた θ に従って切出装置1320は、先に述べた切出ステップを行い、フォーカス・イメージを切り出す。そして、表示装置1140に出力し、ユーザに命令されれば、切り出されたフォーカス・イメージ内に属するデータをデータベース1300から取り出し、出力デバイス1330に引き渡す。出力デバイス1330は、適当な形式でユーザ所望のデータを出力する。また、ユーザは、例えば入力デバイス1360からコンフィデンス最大化ルールを解くように命じ、最小限度のサポートを入力する。すると、制御装置1340は先に示した処理Bを行うように、条件 θ を設定し、切出装置1320に出力する。そして、命じられたコンフィデンス最大化ルールに合致するような領域を解くべく、条件 θ を変化させる等の処理を行う。先に述べたサポート最大化ルール（処理C）や、最適化エントロピー・ルール（処理D）、最適化インタクラスバリエーション・ルール（処理E）、その他凸包上に位置する領域を切出す処理Fに適した条件 θ を切出装置1320に渡す処理を制御装置1340は行う。ユーザは入力デバイス1360から処理の種類や、先に述べたような条件（ θ のみならず、minconf、minsupも）を入力する。また、制御装置1340は、上述の処理Gを行うために平面構成装置1310に、切り出したフォーカス・イメージの $V(i, j)$ を除去する等の処理を命じる。

【0103】以上、本発明を特別の装置にする一例を示したが、本発明はこれに限定されるものではない。例えば、切出装置1320の出力は、出力制御装置を介して出力デバイス1160及び表示装置1140に出力されるようにしてもよいし、この場合出力制御装置からデー

データベースを参照してデータを取り出すようにしてもよい。

【0104】以上は、通常データが有するK個の数値属性のうち2項を選択し、それらの数値属性間の壮観を見つucker処理であったが、数23を目的関数とし、n次元空間の領域を切り出すことができれば、n次元の探索に拡張することができる。

【0105】

【効果】2項以上の数値属性と真偽をとる属性を有するデータ間の結合ルールを見い出すための一手法を提供することができた。

【0106】また、データ間の結合ルールを人間がより把握しやすい形で提示することもできた。そして、多くの結合ルールを可視化することにより、使用する人間の選択の幅を増大させ、より重要な結合ルールを見い出すこと可能とすることもできた。

【0107】さらに、(1) サポート最大化ルールや、(2) コンフィデンス最大化ルール、(3) 最適化エントロピー・ルール、(4) 最適化インタクラスバリエーション・ルールを満たすような範囲(領域)を導出可能とすることもできた。

【0108】また、上記のようなデータ間の結合ルールを高速に実行できるような手法を提供することもできた。

【0109】例えば、ある割合以上で、例えばアウトドアスポーツに興味を示す(真偽をとる属性に相当する)、できるだけまとまった領域に入る顧客を知ることができるので、その条件に合致する多くの顧客に知ってもらいたいダイレクトメールの宛て先を知るのに用いることができる。(サポート最大化ルール)

【0110】一定数以上の顧客を含む、例えば定期預金残高200万円以上の顧客割合が最も高いところを知ることができるので、顧客を絞りこみつつ、有効な宣伝活動を行うことができる。(コンフィデンス最大化ルール)

【図面の簡単な説明】

【図1】平面構成ステップのフローを示す図である。

【図2】領域切り出しステップのための前準備のフローを示す図である。

【図3】直交凸領域を説明するための図である。

【図4】U(S)、V(S)平面の説明をするための図である。

【図5】領域切り出しステップにおける表記を説明するための図である。

【図6】第m-1列から第m列への変化傾向を説明するための図である。

【図7】直交凸領域の各列の状態遷移を表す図である。

【図8】W-Typeにおける第m列と第m-1列の関係を説明するため図であって、(a)は数30の(2)式、(b)は(3)式、(c)は(4)式を説明するた

めの図である。

【図9】 $f_{\alpha}^n(s, t)$ の計算順序を示した図である。

【図10】U-Typeの計算に用いる前処理のアルゴリズムを説明するための図である。

【図11】U-Typeにおける第m列と第m-1列の関係を説明するための図であって、(a)は数33の(2)式、(b)は(3)式、(c)は(4)式を説明するための図である。

【図12】 $f_{\alpha}^n(s, t)$ の計算順序を示した図である。

【図13】D-Typeの計算に用いる前処理のアルゴリズムを説明するための図である。

【図14】D-Typeにおける第m列と第m-1列の関係を説明するための図であって、(a)は数36の(2)式、(b)は(3)式、(c)は(4)式を説明するための図である。

【図15】 $f_{\alpha}^n(s, t)$ の計算順序を示した図である。

【図16】N-Typeにおける第m列と第m-1列の関係を説明するための図であって、(a)は数37の(1)式、(b)は(3)式、(c)は(4)式、(d)は(5)式、(e)は(6)式、(f)は(7)式を説明するための図である。

【図17】 $f_{\alpha}^n(s, t)$ の計算順序を示した図である。

【図18】 $H^x(m, [s, t])$ のデータ構造を示すための図である。

【図19】領域切り出しステップの処理フローを表す図である。

【図20】複数のフォーカス・イメージを見つけ出す処理のフローを示す図である。

【図21】コンフィデンス最大化ルールを導出するための処理の一部を示すための図である。

【図22】コンフィデンス最大化ルールを導出するための処理の一部を示すための図である。

【図23】サポート最大化ルールの導出するための処理の一部を示すための図である。

【図24】サポート最大化ルールの導出するための処理の一部を示すための図である。

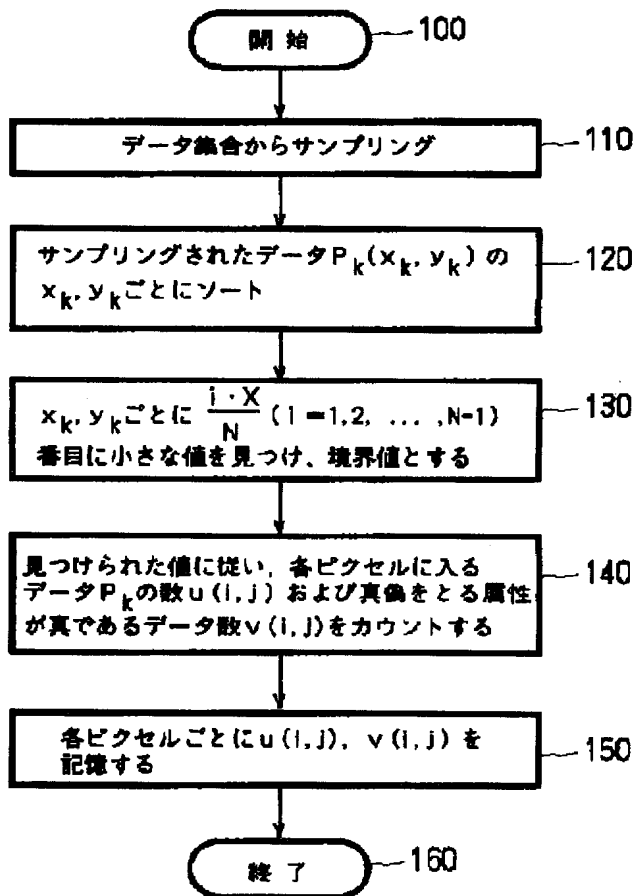
【図25】通常のコンピュータ・システムで本発明を実施した場合の装置構成の一例を示す図である。

【図26】本発明を専用の装置で実施した場合のブロック図である。

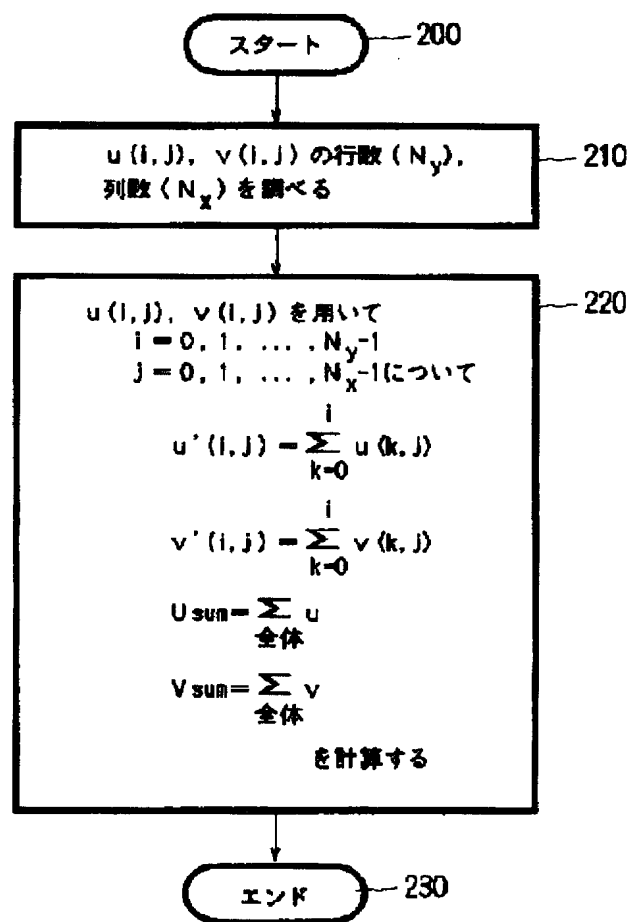
【符号の説明】

1010	CPU	1020	メインメモリ
1030	FDD	1040	CD-ROMドライブ
1050	HDD	1060	表示装置
1070	入力デバイス		
1310	平面構成装置		
1300	データベース	1320	切出装置
1350	表示装置	1130	入力デバイス
1330	出力デバイス	1340	制御デバイス

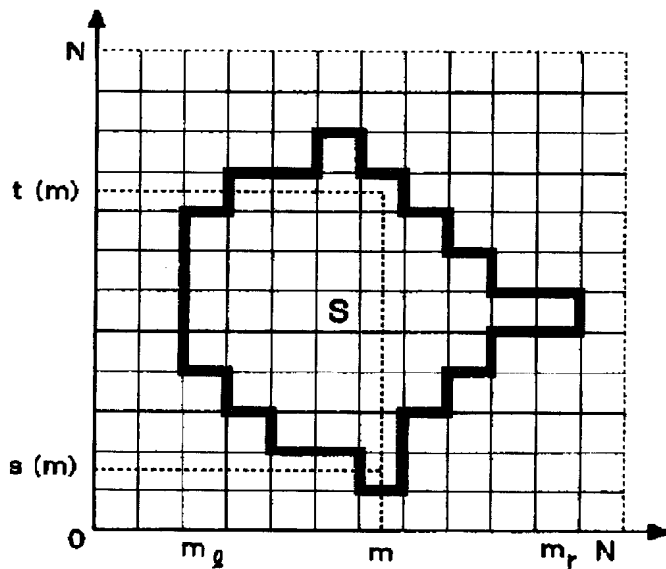
【図1】



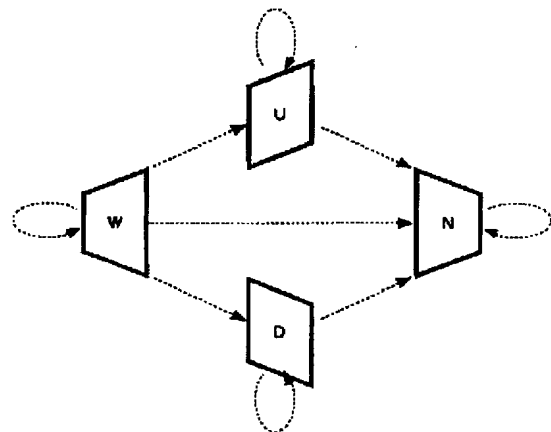
【図2】



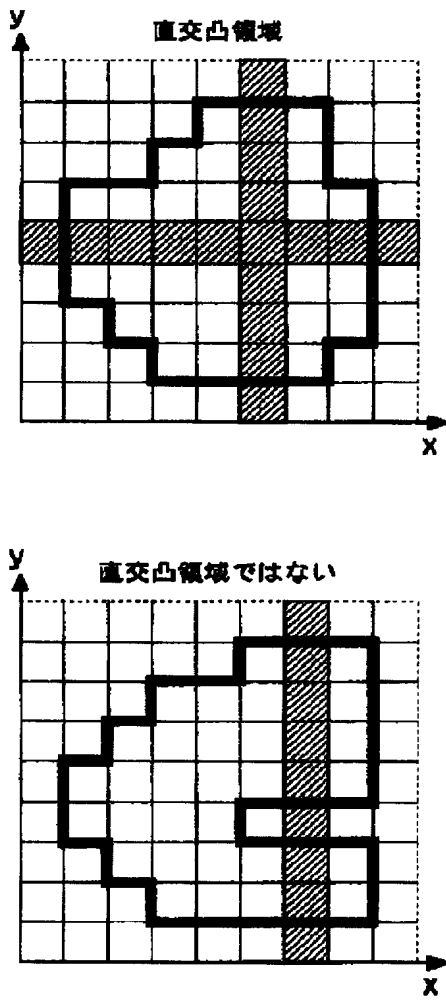
【図5】



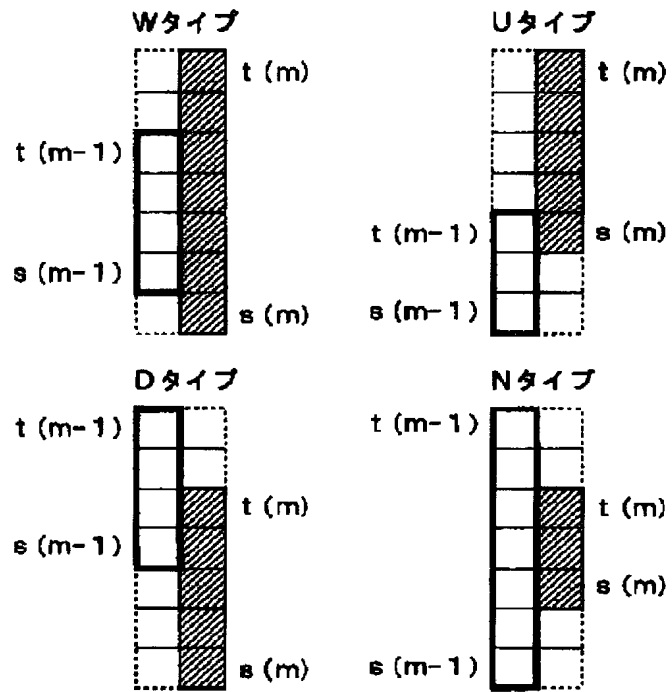
【図7】



【図3】



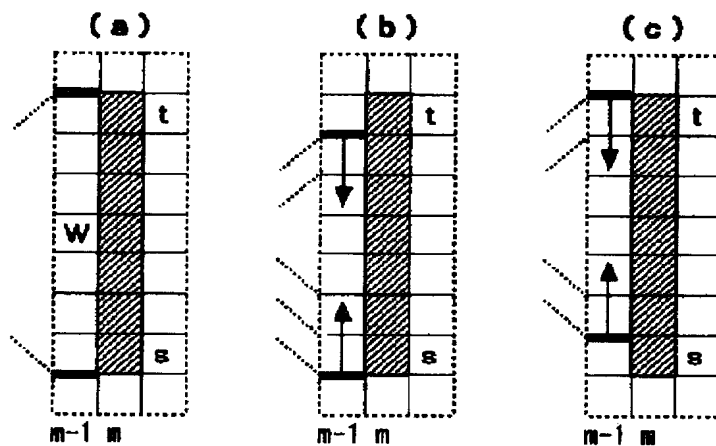
【図6】



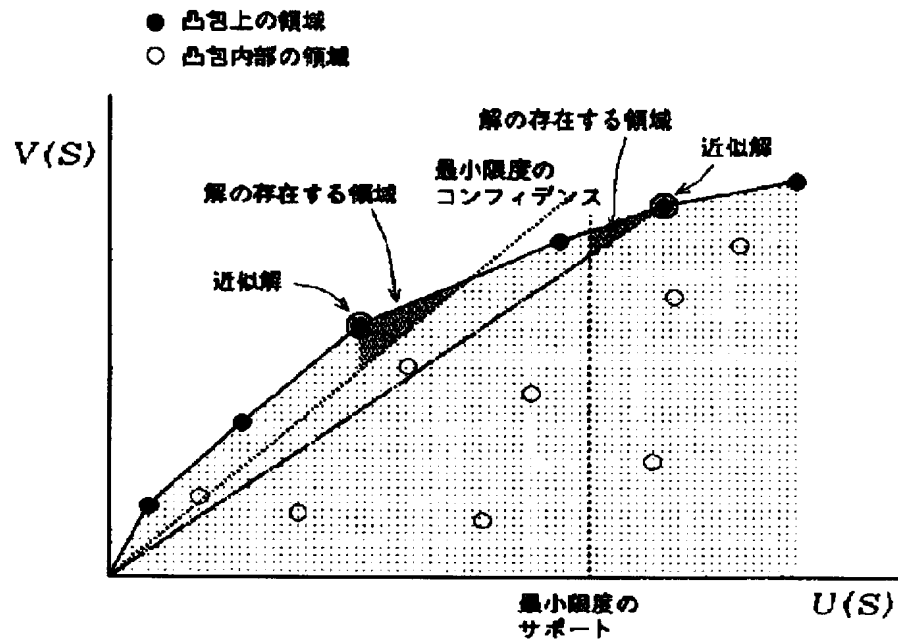
【図18】

31, 30	29, ..., 0
変化傾向 Y	区間 [u, v]

【図8】



【図4】



【図9】

```

for k=0 to  $N_y-1$  do
  for s=0 to  $N_y-1-k$  do
    t=s+k
     $f_n^W(s, t)$  を数30を用いて計算
  end
end
end

```

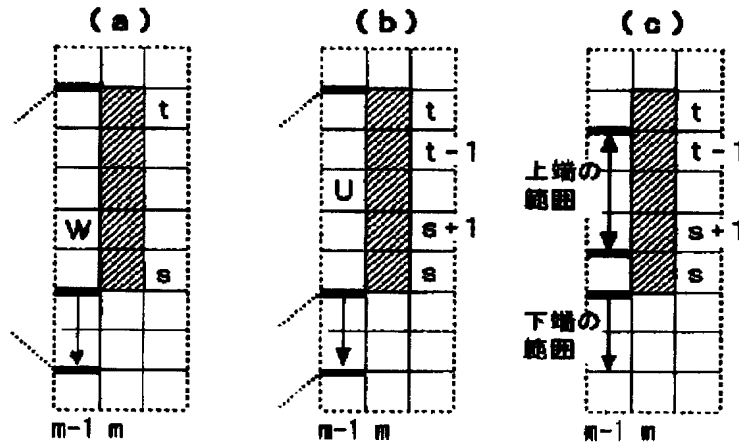
【図10】

```

for t:=0 to  $N_y-1$  do
   $\beta^W=0$ 
   $\beta^U=0$ 
  for s:=0 to t do
    if  $f_{m-1}^W(s, t) > f_{m-1}^W(\beta^W, t)$  then  $\beta^W=s$ 
    if  $f_{m-1}^U(s, t) > f_{m-1}^U(\beta^U, t)$  then  $\beta^U=s$ 
     $\beta_{m-1}^W(s, t)=\beta^W$ 
     $\beta_{m-1}^U(s, t)=\beta^U$ 
  end
end
end

```

【図11】



【図12】

```

for s=0 to  $N_y-1$  do
  for t=s to  $N_y-1$  do
     $f_m^U(s, t)$  を数33を用いて計算
  end
end
end

```

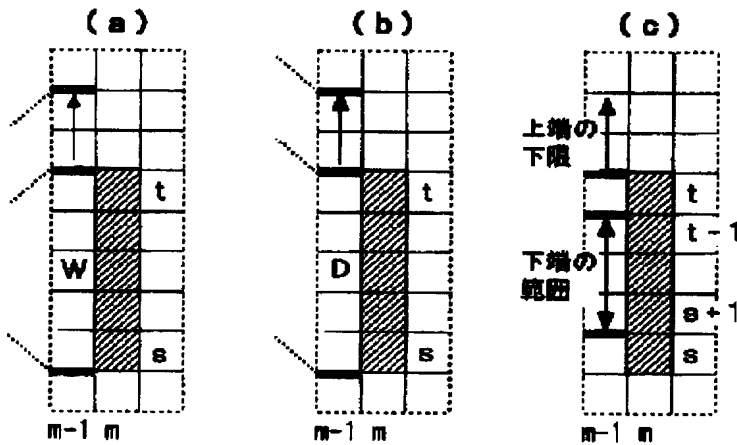
【図13】

```

for s:= $N_y-1$  to 0 do
   $\tau^W = N_y-1$ 
   $\tau^D = N_y-1$ 
  for t:= $N_y-1$  to s do
    if  $f_{m-1}^W(s, t) > f_{m-1}^W(s, \tau^W)$  then  $\tau^W = t$ 
    if  $f_{m-1}^D(s, t) > f_{m-1}^D(s, \tau^D)$  then  $\tau^D = t$ 
     $\tau_{m-1}^W(s, t) = \tau^W$ 
     $\tau_{m-1}^D(s, t) = \tau^D$ 
  end
end
end

```


【図14】



【図15】

```

for t=Ny-1 to 0 do
  for s=t to 0 do
     $f_m^D(s, t)$  を数36を用いて計算
  end
end
end

```

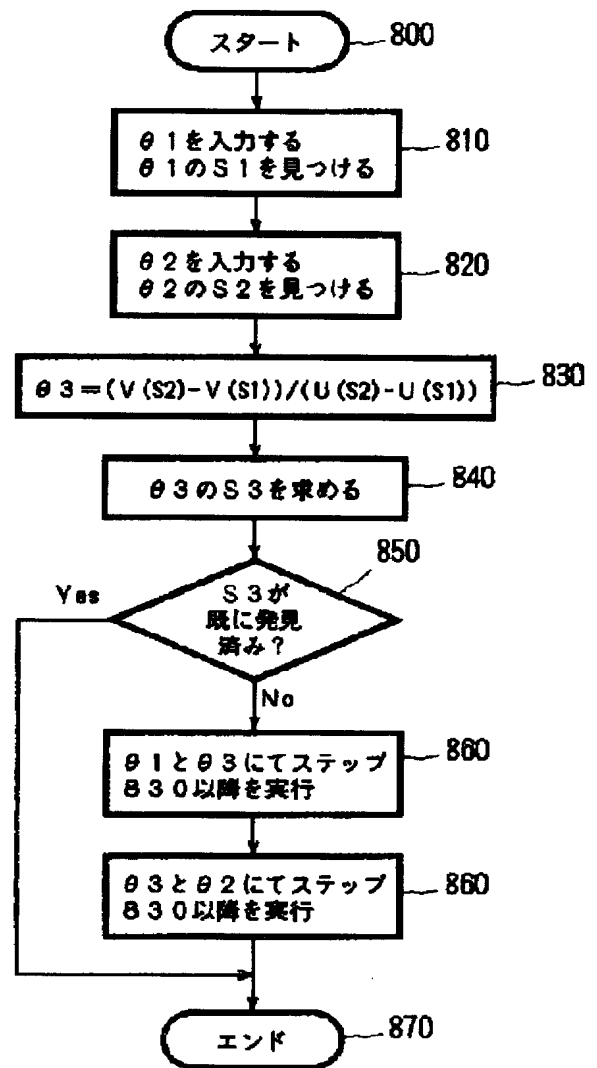
【図17】

```

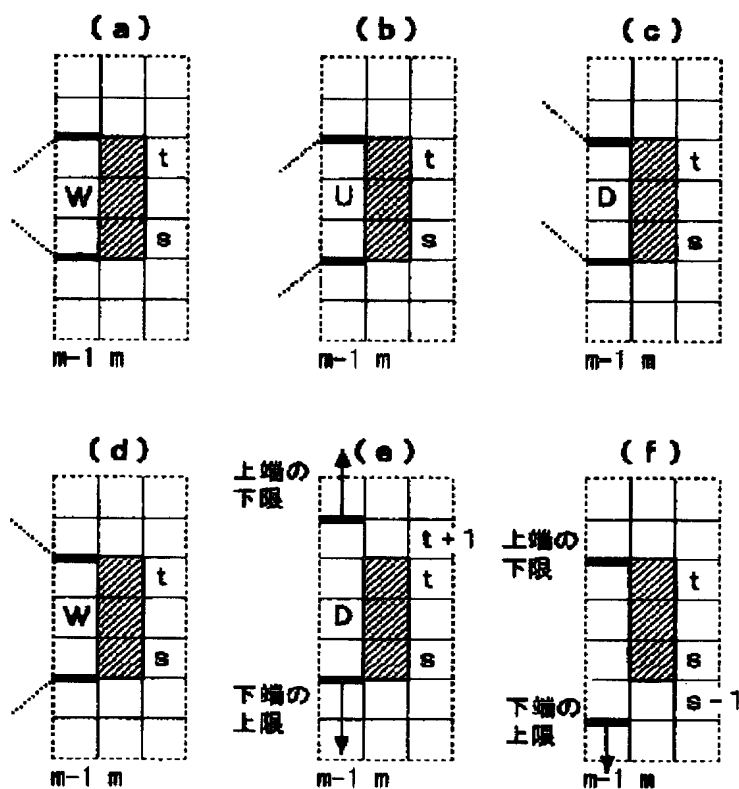
for k=Ny-1 to 0 do
  for t=Ny-1 to k do
    s=t-k
     $f_m^N(s, t)$  を数37を用いて計算
  end
end
end

```

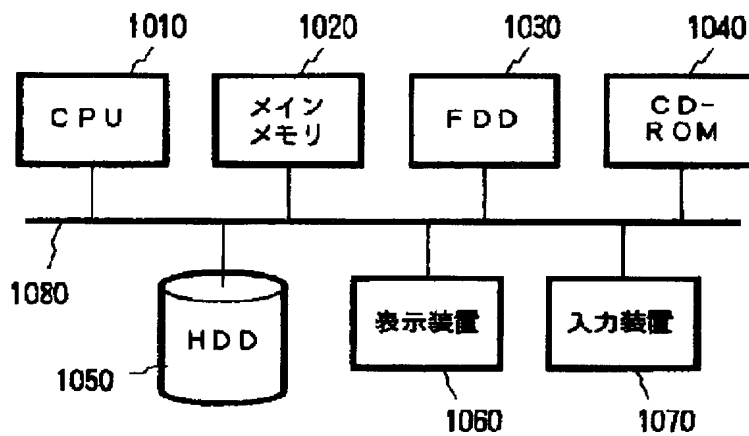
【図20】



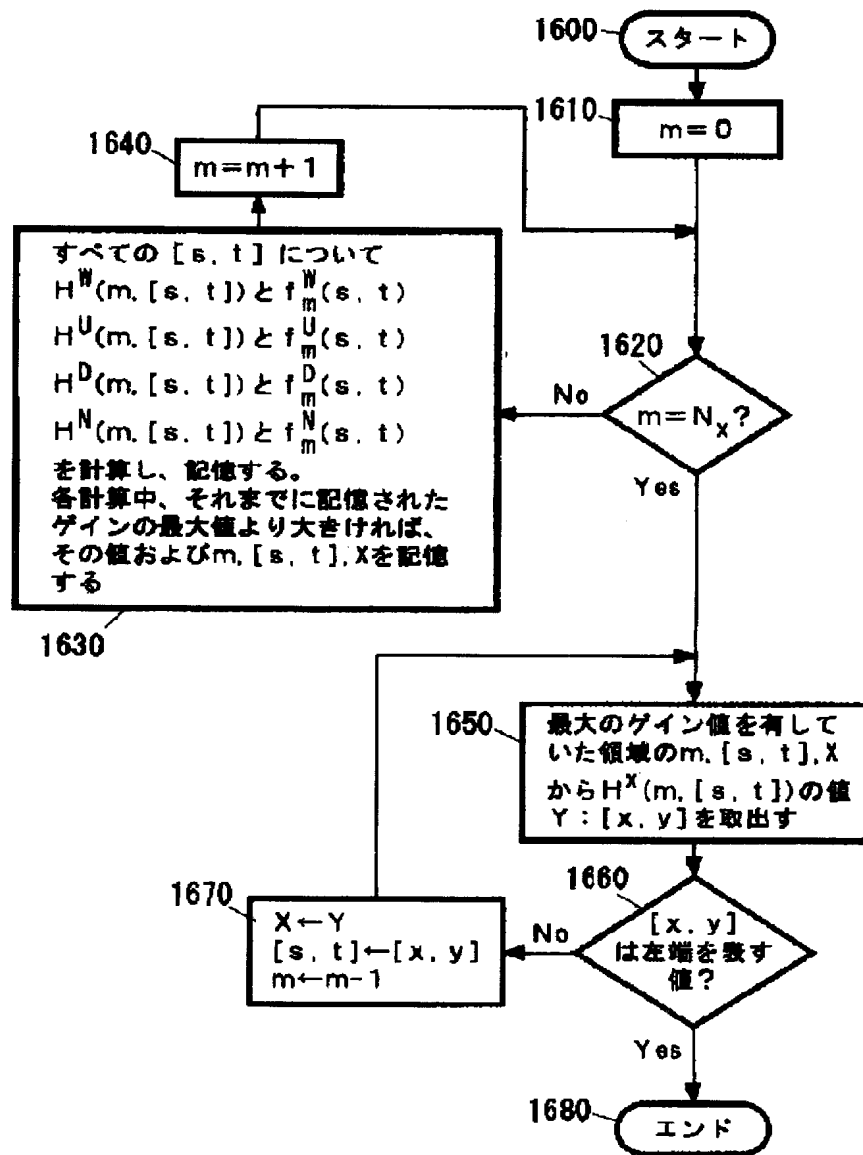
【図16】



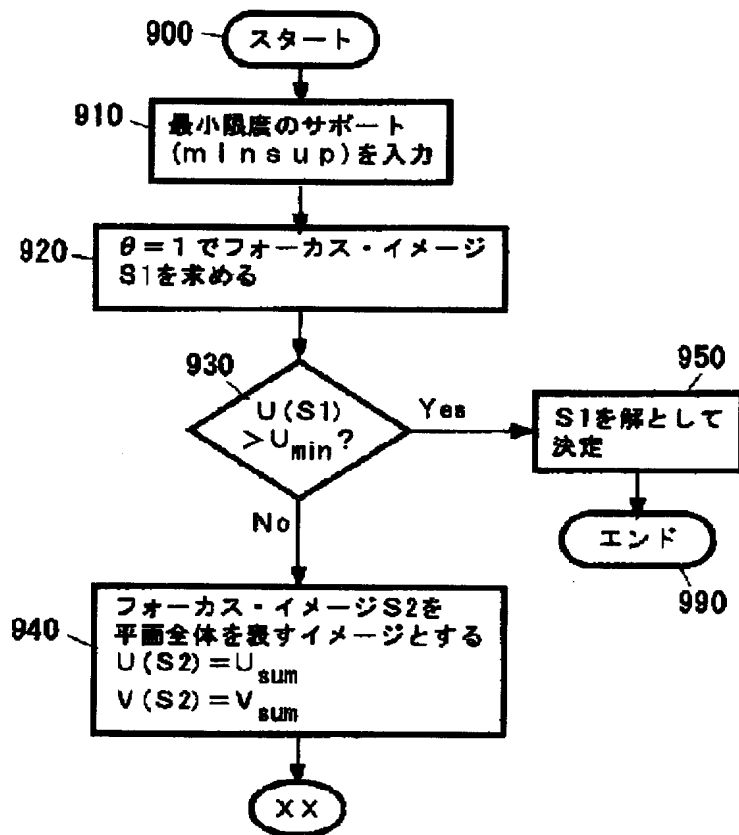
【図25】



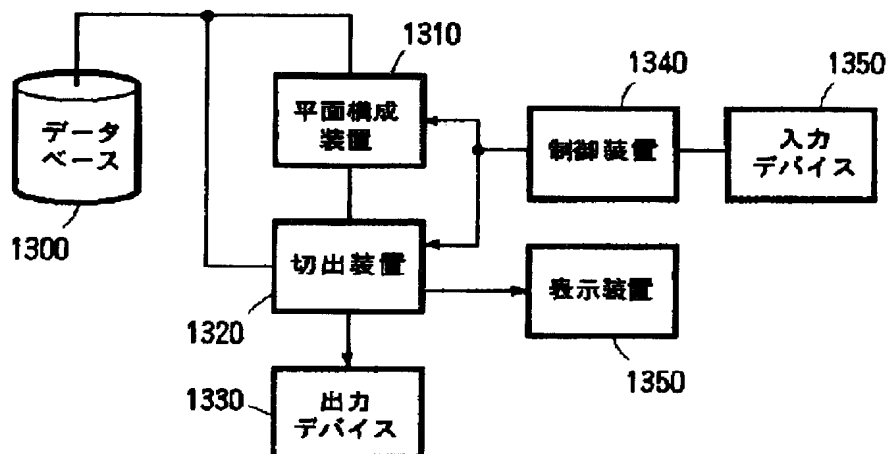
【図19】



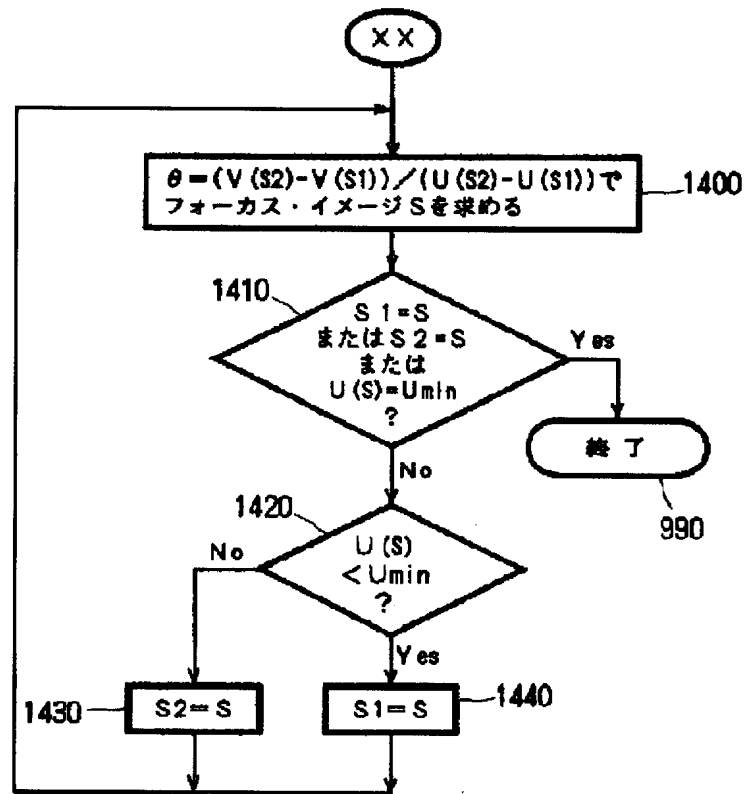
【図21】



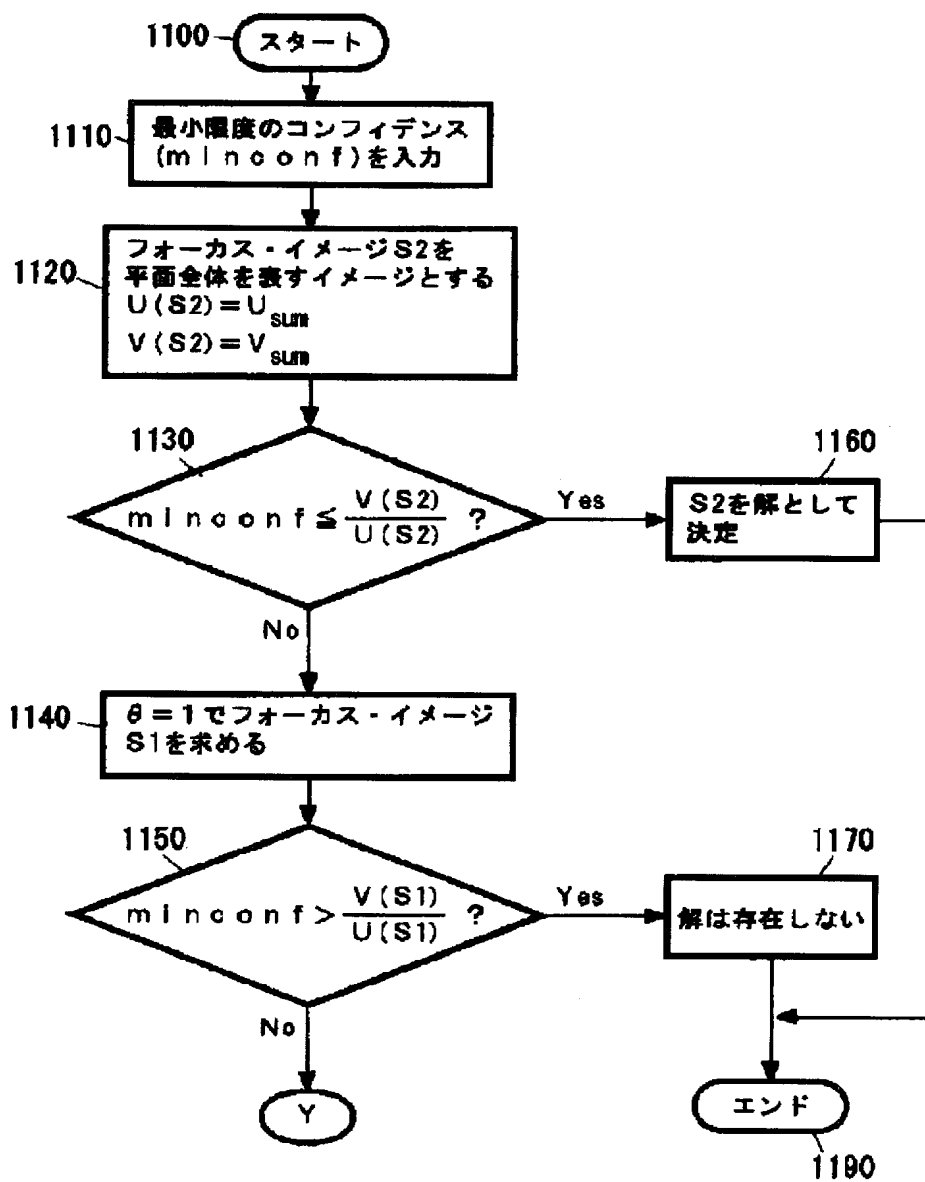
【図26】



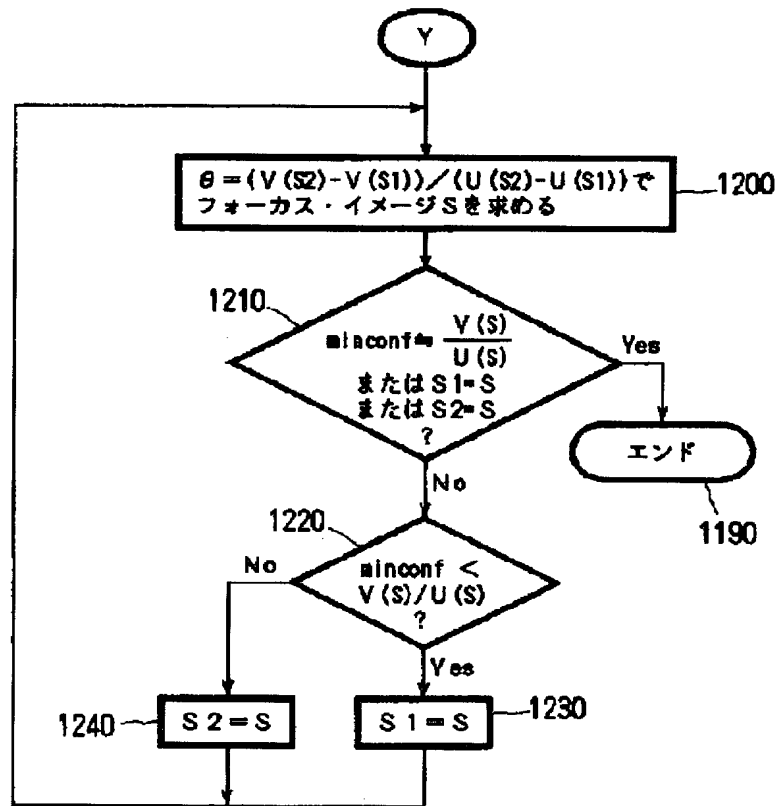
【図 22】



【図23】



【図24】



【手続補正書】

【提出日】平成9年10月7日

【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】0018

【補正方法】変更

【補正内容】

【0018】さらに、切り出された直交凸領域S内の各ピクセルの $v(i, j)$ 、 $u(i, j)$ が、平面全体のデータ数に対する平面全体の真偽をとる属性が真であるデータ数の割合に等しくなるよう $v(i, j)$ を変更し、当該変更された $v(i, j)$ を用いて、入力された条件 θ_4 に従い、

【数20】

$$\sum_{(i,j) \in S_4} g(i,j) = \sum_{(i,j) \in S_4} (v(i,j) - \theta_4 u(i,j))$$

を最大にするようなピクセルの第4の領域 S_4 を切り出すようにすることも考えられる。このようにすると、二次的な相関ルールを導き出すことができる。

【手続補正2】

【補正対象書類名】明細書

【補正対象項目名】0019

【補正方法】変更

【補正内容】

【0019】また、先の平面構成ステップは、複数のデータから、X個のデータをランダムサンプリングし、サンプリングされたデータを各数値属性についてソートし、 $X \cdot i \div N$ ($i = 1, 2, \dots, N$) 番目に該当する数値及び $X \cdot n \div M$ ($n = 1, 2, \dots, M$) 番目に該当する数値を記憶し、記憶された数値を基準にして、複数のデータを $N \times M$ 個のピクセルに入れるようにすることも考えられる。このようにすると、各行各列にデータを高速にまたほぼ均等に割り振ることができる。

【手続補正3】

【補正対象書類名】明細書

【補正対象項目名】0071

【補正方法】変更

【補正内容】

【0071】次に領域を記憶しておく配列を用意する。これは、 $N_x \times N_y \cdot (N_y + 1) \div 2$ の整数型2次元配列であり、W、U、D、N-Typeのそれぞれに対して1つ用意する。この要素を $H^x(m, [s, t])$ ($0 \leq m \leq N$ 、

$-1, 0 \leq [s, t] \leq N_y (N_y + 1) / 2, X \in \{W, U, D, N\}$) と表すこととする。

【手続補正 4】

【補正対象書類名】明細書

【補正対象項目名】0 1 0 4

【補正方法】変更

* 【補正内容】

【0 1 0 4】以上は、通常データが有する K 個の数値属性のうち 2 項を選択し、それらの数値属性間の相関を見つける処理であったが、数 2 3 を目的関数とし、 n 次元空間の領域を切り出すことができれば、 n 次元の探索に拡張することができる。

*

フロントページの続き

(72)発明者 福田 剛志

神奈川県大和市下鶴間1623番地14 日本アイ・ビー・エム株式会社 東京基礎研究所内

(72)発明者 徳山 豪

神奈川県大和市下鶴間1623番地14 日本アイ・ビー・エム株式会社 東京基礎研究所内

(72)発明者 森下 真一

神奈川県大和市下鶴間1623番地14 日本アイ・ビー・エム株式会社 東京基礎研究所内